Gaze Guidance, Task-Based Eye Movement Prediction, and Real-World Task Inference using Eye Tracking

Srinivas Sridharan
sxs9716@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Dissertation is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritscholarworks@rit.edu.
Gaze Guidance, Task-Based Eye Movement Prediction, and Real-World Task Inference using Eye Tracking

by

Srinivas Sridharan

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computing and Information Sciences

PhD Program in Computing and Information Sciences
B. Thomas Golisano College of Computing and Information Sciences

Rochester Institute of Technology
Rochester, New York
July 22 2016
Gaze Guidance, Task-Based Eye Movement Prediction, and Real-World Task Inference using Eye Tracking

by

Srinivas Sridharan

Committee Approval:
We, the undersigned committee members, certify that we have advised and/or supervised the candidate on the work described in this dissertation. We further certify that we have reviewed the dissertation manuscript and approve it in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Computing and Information Sciences.

Dr. Reynold Bailey (Dissertation Advisor)  Date
Associate Professor

Dr. Joe Geigel  Date
Professor

Dr. Anne Haake  Date
Professor and Dean

Dr. Linwei Wang  Date
Associate Professor

Dr. Daniel Phillips (Dissertation Defense Chair)  Date
Professor

Certified by:

Dr. Pengcheng Shi  Date
Director, Ph.D. program Computing and Information Sciences
Gaze Guidance, Task-Based Eye Movement Prediction, and Real-World Task Inference using Eye Tracking

by

Srinivas Sridharan

Abstract

The ability to predict and guide viewer attention has important applications in computer graphics, image understanding, object detection, visual search and training. Human eye movements provide insight into the cognitive processes involved in task performance and there has been extensive research on what factors guide viewer attention in a scene. It has been shown, for example, that saliency in the image, scene context, and task at hand play significant roles in guiding attention.

This dissertation presents and discusses research on visual attention with specific focus on the use of subtle visual cues to guide viewer gaze and the development of algorithms to predict the distribution of gaze about a scene. Specific contributions of this work include: a framework for gaze guidance to enable problem solving and spatial learning, a novel algorithm for task-based eye movement prediction, and a system for real-world task inference using eye tracking.

A gaze guidance approach is presented that combines eye tracking with subtle image-space modulations to guide viewer gaze about a scene. Several experiments were conducted using this approach to examine its impact on short-term spatial information recall, task sequencing, training, and password recollection. A model of human visual attention prediction that uses saliency maps, scene feature maps and task-based eye movements to predict regions of interest was also developed. This model was used to automatically select target regions for active gaze guidance to improve search task performance. Finally, we develop a framework for inferring real-world tasks using image features and eye movement data.

Overall, this dissertation naturally leads to an overarching framework, that combines all three contributions to provide a continuous feedback system to improve performance on repeated visual search tasks. This research has important applications in data visualization, problem solving, training, and online education.
Acknowledgements

I would like to express the deepest appreciation to my advisor and mentor, Dr. Reynold Bailey, for the support, encouragement, inspiration, and guidance he provided me throughout my graduate study. His constant guidance has shaped me into an independent researcher with appreciation for this field and motivated me to explore and solve new challenges in my area of research.

I am grateful to my dissertation committee members: Dr. Joe Geigel, Dr. Anne Haake, and Dr. Linwei Wang for their insightful suggestions and comments. I would like to thank Dr. Ray Ptucha, Dr. Chris Kanan, and Dr. Gabriel Diaz for their guidance and feedback. I would also like to thank my collaborators Dr. Ann McNamara and Dr. Cindy Grimm for their support on several peer-reviewed conference publications.

My appreciation goes to my colleagues and good friends from the Computer Graphics and Applied Perception Lab, Tom Booth, Vasudev Bethamcherla, Brendan John, Justin Bennet, James Pieszala, Amog Rajenderan, Junxia Xu, Yuqiong Wang, Philip Van Der Lugt, James Coddington, Darrel Pollard, and Michael Fink for the useful and stimulating discussions.

My sincere thanks also go to Ph.D. students Zack Fitzsimmons, Jingjia Xu, David Narvaez, Xuan Guo, and Matthew Le for all the thought provoking discussions and conversations.

From the bottom of my heart, I would like to thank all the eye tracking experiment volunteers who provided their valuable time for science.

Lastly, I would like to express my gratitude to my dear parents, my wife, and my brother for their support, love, presence, and patience.

The work presented in this dissertation is supported by National Science Foundation (NSF) under Award No. IIS-0952631. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.
Contents

Abstract iii

Acknowledgements iv

List of Figures viii

List of Tables x

1 Introduction 1
   1.1 Overview ................................................................. 1
   1.2 Organization of Dissertation ....................................... 4

2 General Background 5
   2.1 Human Visual System ................................................. 5
   2.2 Visual Attention ..................................................... 8
       2.2.1 Bottom-up Saliency-Based Visual Attention ................. 9
       2.2.2 Top-Down Cognition-Based Visual Attention ............... 10
       2.2.3 Computational Visual Attention Models ..................... 10
   2.3 Eye Tracking ......................................................... 13

3 Gaze Guidance for Problem Solving and Spatial Learning 16
   3.1 Adaptive Subtle Gaze Direction Technique ........................ 19
       3.1.1 Experiment Design ............................................. 21
       3.1.2 Results and Discussion ....................................... 22
       3.1.3 Conclusion ....................................................... 23
   3.2 Impact of Subtle Gaze Direction on Short-Term Spatial Information Recall 23
       3.2.1 Experiment Design ............................................. 24
           3.2.1.1 Stimuli .................................................. 25
           3.2.1.2 Participants ............................................ 25
           3.2.1.3 Procedure ............................................... 27
       3.2.2 Results and Discussion ....................................... 29
           3.2.2.1 Counting Error .......................................... 30
           3.2.2.2 Location Error ......................................... 31
           3.2.2.3 Shape Error ............................................ 34
           3.2.2.4 Percentage Gaze Time .................................. 34
3.2.3 Conclusion ........................................... 37
3.3 Directing Gaze in Narrative Art .......................... 37
  3.3.1 Experiment Design .................................. 39
  3.3.2 Stimuli ............................................. 40
    3.3.2.1 Analysis of Data ............................... 43
  3.3.3 Results and Discussion .............................. 44
  3.3.4 Conclusion ......................................... 48
3.4 Subtle Gaze Manipulation for Improved Mammography Training .......................... 49
  3.4.1 Experiment Design .................................. 51
    3.4.1.1 Stimuli ......................................... 51
    3.4.1.2 Participant ..................................... 53
    3.4.1.3 Procedure ...................................... 54
  3.4.2 Results and Discussion .................................. 57
    3.4.2.1 Training Session Results ....................... 59
  3.4.3 Short-Term Follow-up Session Results ................. 61
    3.4.3.1 Long-Term Follow-up Session Results ............. 63
    3.4.3.2 Receiver Operating Characteristic (ROC) Results .... 64
  3.4.4 Conclusion ......................................... 66
3.5 Gaze Guidance for Improved Password Recollection .......................... 67
  3.5.1 Experiment Design .................................. 69
    3.5.1.1 Stimuli ......................................... 70
    3.5.1.2 Participants .................................... 70
    3.5.1.3 Procedure ...................................... 71
    3.5.1.4 Memorization Stage ............................. 71
    3.5.1.5 Cognitive Task ................................. 71
    3.5.1.6 Recall Stage .................................... 72
  3.5.2 Results and Discussion .................................. 73
    3.5.2.1 Accuracy ........................................ 73
    3.5.2.2 Response Time ................................... 75
  3.5.3 Dwell-time on Target Regions ......................... 76
  3.5.4 Conclusion ......................................... 76
3.6 Guiding Attention in Controlled Real-World Environments ......................... 77
  3.6.1 Experiment Design .................................. 79
    3.6.1.1 System Overview ................................. 80
    3.6.1.2 SIFT descriptor extraction and matching .......... 82
    3.6.1.3 Stability Considerations ....................... 84
    3.6.1.4 User Study ..................................... 84
  3.6.2 Results and Discussion .................................. 86
  3.6.3 Conclusion ......................................... 89
4 Task-Based Eye Movement Prediction ........................................ 91
  4.1 Automatic Target Prediction and Subtle Gaze Guidance for Improved Spatial Information Recall .......................... 92
    4.1.1 Automatic Target Prediction Framework ................. 94
    4.1.2 Experiment Design .................................. 97
      4.1.2.1 Stimuli ......................................... 98
      4.1.2.2 Participants .................................... 98
4.1.2.3 Procedure ................................. 99
4.1.3 Results and Discussion ................................. 101
  4.1.3.1 Counting Error ..................................... 102
  4.1.3.2 Location Error ..................................... 104
  4.1.3.3 Shape Error ...................................... 106
  4.1.3.4 Percentage Gaze Time .................................. 108
4.1.4 Conclusion ....................................... 110

5 Real-World Task Inference using Eye Tracking 112
  5.1 Real-World Task Inference using Visual Bag-Of-Features and Eye-Movement Data ............................................. 113
    5.1.1 System Design ......................................... 115
      5.1.1.1 Visual Bag-Of-Features Extraction .................. 116
      5.1.1.2 Eye-Movements Feature Extraction ..................... 118
      5.1.1.3 Combining Visual Bag-Of-Features and Eye-Movement Features ............................................. 119
    5.1.2 Experiment Design ..................................... 120
      5.1.2.1 Participants ........................................ 120
      5.1.2.2 Tasks ............................................. 120
      5.1.2.3 Procedure ........................................ 122
      5.1.2.4 Evaluation ........................................ 123
    5.1.3 Results and Discussion .................................. 124
      5.1.3.1 Results From Bag-Of-Features ...................... 124
      5.1.3.2 Results From Eye-Movement Features ................ 128
    5.1.4 Results From Eye-Movement Features + Bag-of-Features ................................. 132
    5.1.5 Conclusion ........................................ 135

6 Conclusion and Future Work 136
  6.1 Future Work ........................................ 137

A Eye Tracking Datasets 141

Bibliography 143
## List of Figures

1.1 Illustration of gaze guidance, visual attention prediction, and task inference. 2

2.1 Schematic diagram of the human visual system 6
2.2 Distribution of cones in the retina 7
2.3 Yarbus eye tracking experiment 1967 11
2.4 Taxonomy of visual attention studies 12
2.5 Pupil center and corneal reflections 14
2.6 Remote and head mounted eye tracking systems 15

3.1 Subtle Gaze Direction overview 17
3.2 Gaussian modulation for adaptive subtle gaze direction 20
3.3 Gaze distribution for an image under static and modulated conditions 22
3.4 Example solution from impact of SGD on spatial information recall experiment 28
3.5 Average counting error for spatial information recall 30
3.6 Average counting error by target regions for spatial information recall 31
3.7 Average location error for spatial information recall 32
3.8 Average location error by target regions for spatial information recall 33
3.9 Outlier image related to location error for spatial information recall 33
3.10 Average shape error for spatial information recall 35
3.11 Percentage gaze time on target regions for spatial information recall 35
3.12 “The Tribute Money” - by Masaccio 38
3.13 An example narrative art image with panels highlighted 42
3.14 Normalized Levenshtein distance measure between groups 45
3.15 Normalized Levenshtein distance measure between groups per image ranked on number of panels 45
3.16 Scanpath and heat-map comparison between groups 47
3.17 Comparison of simplified-expert vs. adversarial scanpaths on a mammogram image 55
3.18 Illustration of hits, close matches, and misses for accuracy 58
3.19 Average weight-based accuracy during training 60
3.20 Binary classification-based accuracy during training 60
3.21 Average weight-based accuracy during short-term follow-up session 61
3.22 Binary classification-based accuracy during short-term follow-up session 62
3.23 Average weight-based accuracy during long-term follow-up session 63
3.24 Binary classification-based accuracy during long-term follow-up session 64
3.25 Sensitivity vs. specificity across all sessions 64
3.26 ROC plots across all sessions 65
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.27</td>
<td>Overall experiment design for password recollection</td>
<td>68</td>
</tr>
<tr>
<td>3.28</td>
<td>Accuracy of correct password digits recalled</td>
<td>73</td>
</tr>
<tr>
<td>3.29</td>
<td>Accuracy of correct but misplaced password digits recalled</td>
<td>74</td>
</tr>
<tr>
<td>3.30</td>
<td>Response time for password recall</td>
<td>75</td>
</tr>
<tr>
<td>3.31</td>
<td>Scanpath comparison between gaze guided approaches</td>
<td>76</td>
</tr>
<tr>
<td>3.32</td>
<td>Real world subtle gaze manipulation experiment setup</td>
<td>77</td>
</tr>
<tr>
<td>3.33</td>
<td>System architecture and data flow for real world subtle gaze guidance</td>
<td>81</td>
</tr>
<tr>
<td>3.34</td>
<td>SIFT descriptor extraction and matching</td>
<td>83</td>
</tr>
<tr>
<td>3.35</td>
<td>Levenshtein distances across all sequences for all participants</td>
<td>87</td>
</tr>
<tr>
<td>3.36</td>
<td>Histogram of Levenshtein distances across all sequences for all participants</td>
<td>87</td>
</tr>
<tr>
<td>3.37</td>
<td>Gaze-guided scanpath for a single subject</td>
<td>88</td>
</tr>
<tr>
<td>3.38</td>
<td>Subjects performance on parts retrieval task</td>
<td>89</td>
</tr>
<tr>
<td>4.1</td>
<td>Overall system design for automatic target prediction model</td>
<td>95</td>
</tr>
<tr>
<td>4.2</td>
<td>Automatic target prediction: visualization showing Gaussian pulse with edge modulation</td>
<td>97</td>
</tr>
<tr>
<td>4.3</td>
<td>Example solution from one subject on spatial information recall using automatic targets</td>
<td>100</td>
</tr>
<tr>
<td>4.4</td>
<td>Average counting error for spatial information recall using automatic targets</td>
<td>102</td>
</tr>
<tr>
<td>4.5</td>
<td>Average counting error as number of target regions for spatial information recall using automatic targets</td>
<td>104</td>
</tr>
<tr>
<td>4.6</td>
<td>Average location error for spatial information recall using automatic targets</td>
<td>105</td>
</tr>
<tr>
<td>4.7</td>
<td>Average location error as number of target regions for spatial information recall using automatic targets</td>
<td>106</td>
</tr>
<tr>
<td>4.8</td>
<td>Average shape error for spatial information recall using automatic targets</td>
<td>107</td>
</tr>
<tr>
<td>4.9</td>
<td>Average shape error as number of target regions for spatial information recall using automatic targets</td>
<td>108</td>
</tr>
<tr>
<td>4.10</td>
<td>Percentage gaze time spent within the target regions comparison for spatial information recall using automatic targets</td>
<td>109</td>
</tr>
<tr>
<td>5.1</td>
<td>Overall system design for real-world task inference</td>
<td>114</td>
</tr>
<tr>
<td>5.2</td>
<td>Spatial pyramid of SURF features to extract visual bag-of-features</td>
<td>117</td>
</tr>
<tr>
<td>5.3</td>
<td>Real-world tasks used to test the task inference system</td>
<td>121</td>
</tr>
<tr>
<td>5.4</td>
<td>Subset of training patches for each task for classification</td>
<td>125</td>
</tr>
<tr>
<td>5.5</td>
<td>Confusion matrix for the training data-set using bag-of-features classifier</td>
<td>126</td>
</tr>
<tr>
<td>5.6</td>
<td>Confusion matrix for the testing data-set using bag-of-features classifier</td>
<td>127</td>
</tr>
<tr>
<td>5.7</td>
<td>Confusion matrix for the training data-set using eye-movement features</td>
<td>128</td>
</tr>
<tr>
<td>5.8</td>
<td>Confusion matrix for the testing data-set using eye-movement features</td>
<td>129</td>
</tr>
<tr>
<td>5.9</td>
<td>Scanpath image of a subject while washing hands and making coffee</td>
<td>130</td>
</tr>
<tr>
<td>5.10</td>
<td>Scanpath image of a subject while counting chairs and writing name</td>
<td>130</td>
</tr>
<tr>
<td>5.11</td>
<td>Eye movement feature testing results</td>
<td>131</td>
</tr>
<tr>
<td>5.12</td>
<td>Confusion matrix for the training data-set using combined eye-movement features and bag-of-features</td>
<td>133</td>
</tr>
<tr>
<td>5.13</td>
<td>Bag-of-features and eye-movement feature testing</td>
<td>134</td>
</tr>
<tr>
<td>6.1</td>
<td>Gaze-guidance, visual attention prediction, and task inference applied to repeated real-world image search problems</td>
<td>138</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Correct predictions on intended panel sequences for narrative art images . 46
3.2 Mapping of test outcomes to binary classification metrics . . . . . . . . . 58
3.3 Significant t-test results between groups for password recall . . . . . . . 75

5.1 Table showing binary classification mapping of trained features . . . . . 123
5.2 Table showing binary classification using bag-of-features . . . . . . . . . 127
5.3 Table showing binary classification using eye-movement features . . . . 131
5.4 Table showing binary classification using combined eye-movement fea-
tures and bag-of-features . . . . . . . . . . . . . . . . . . . . . . . . . . . . 133

A.1 Information related to eye tracking datasets . . . . . . . . . . . . . . . . . 141
To my parents, my wife Megha, and my brother
for their endless love and support.
Chapter 1

Introduction

1.1 Overview

Humans perform a wide variety of tasks such as navigation, reading, and interacting with objects in the scene. Task performance primarily depends on the input from the environment, and from memory about the task. Visual attention plays a vital role in performing a task with both speed and accuracy. Attention research has been concerned with understanding the input stimuli, neural mechanism, information processing and memory, and the motor signal involved in task performance. It has also interested researchers to understand what guides human attention about an image. Eye tracking technology has played a significant role in understanding visual attention. Eye movements provide information about the regions attended in an image and gives insight about the underlying cognitive process [1]. Salient content such as, regions with high local contrast, high edge density, and bright colors have been shown to attract viewer attention [2–4]. It has also been shown that the pattern of eye movements is heavily influenced by the viewer’s intent or the assigned task [5–7]. Viewers are also naturally
Figure 1.1: Figure shows a subject’s eye movement data and illustrates gaze guidance, visual attention prediction, and task inference. Orange circles show the sequence of fixations. Directing viewer’s visual attention to a relevant region in the image is termed as gaze guidance. Label a shows the intent to guide viewer attention to the speed gauge using an overt visual cue (red arrow). Label b shows a predicted region in the image where the viewer will look next. The prediction model can also be used to generate an entire sequence of eye movements. Label c shows the task inferred using the image content and eye movement data, which in this case is “looking at rear view mirrors.”

drawn to faces or regions with contextual information [8]. Understanding human task performance using only the scene context, [9–11] or “scene + visual attention” information [12, 13] is an active area of research. The dissertation addresses the following areas of research in applied perception and computational visual attention:

- **Gaze Guidance:** The ability to direct viewer attention (see Figure 1.1) has important applications in computer graphics, data visualization, image analysis, and training. Computer-based gaze guidance techniques can be used to aid spatial learning, visual search task completion, and training. Active gaze guidance can be used to assists people performing repeated critical visual search tasks (e.g. radiologists, TSA agents). Subtle gaze guidance techniques achieve similar outcomes and
have minimal impact on the viewing experience. An important precursor to gaze
guidance is predicting where the viewer will look in the scene, and then guiding
viewer attention to unattended task-relevant regions.

- **Visual Attention Prediction**: The ability to predict when and where someone
will look (see Figure 1.1) in a given scene is a very challenging task. It has been
shown that humans process visual information in a need-based manner. We look
for things that are relevant for the current task and pay less attention to irrelevant
objects in the scene. This dissertation processes new computational models to
predict task-based human gaze behavior.

- **Task Inference**: Eye movements can convey vital information of the sub-processes
involved when performing a task such as checking your rear view mirrors while driv-
ing (see Figure 1.1). Eye movements reveal the shift in attention and a sequence of
eye movements highly relate to the task at hand. The process of differentiating eye
movements as task-based and information-gathering-based is an key challenge in
the eye tracking research community. Furthermore, computationally inferring the
task being performed based on eye movement data will have far reaching implica-
tions for training, learning, and assistive-technologies. This dissertation addresses
the task inference problem by leveraging image features and real-time eye move-
ment data to classify the task being performed by the viewer.

In this dissertation the key contributions are: (1) to develop and evaluate an adaptive
approach to guide viewer attention about a scene that requires no permanent or overt
changes to the scene being viewed, and has minimal impact on viewing experience,
(2) to evaluate a model of human visual attention prediction that uses saliency maps,
scene feature maps and task-based eye movements to predict regions of interest, and (3)
develop a framework for task inference based on scene information and eye movement data. Section 1.2 describes the overall organization of this dissertation.

1.2 Organization of Dissertation

Visual attention research has primarily been focused in areas of neuro-physiology and human psychology. In the past three decades, extensive research on computational visual attention, models on neural mechanism of visual attention, and applications of human visual perception to aid computer graphics and computer vision technologies have been published. However, there is no known framework that combines computational visual attention models with applied perception techniques for actively guiding visual attention, to understand, and aid human task performance. The contributions of this dissertation lays the ground work for such a framework.

In this dissertation a chapter is dedicated to each contribution. Each chapter contains sections that include experiments to evaluate the contribution, and provide detailed discussion of the results obtained. Each section is then summarized with a conclusion describing the merits of the contribution. The remainder of this dissertation is organized as follows: Chapter 2 presents general background and related work, chapter 3 presents gaze guidance for problem solving and spatial learning, chapter 4 presents a model for task-based eye movement prediction, chapter 5 describes a framework for real-world task inference using eye tracking, chapter 6 summarizes all the key contribution and provides the conclusion and future work for each contribution, and the appendix section provides a list of image datasets used in this dissertation.
Chapter 2

General Background

This chapter provides general background information for readers who may not be familiar with the human visual system, visual attention models, and video-based eye tracking technologies. This chapter provides definition of key-terms and concepts to understand the work described in this dissertation.

2.1 Human Visual System

Humans rely heavily on vision to navigate in their environment, and to perform day-to-day tasks with ease. Nearly half of the human brain is directly or indirectly devoted to vision. Figure 2.1 shows the schematic diagram of the horizontal cross-section of a human eye, highlighting the important parts such as the cornea, iris, pupil, lens, retina and fovea. The fovea is located on the retina which is the central focus point of high acuity vision. In humans, acuity at the fovea is higher when compared to periphery. The falloff in visual acuity as distance from the fovea increases, is directly related to the distribution of the cones in the retina [15].
Figure 2.1: Figures shows the diagram of the human visual system. It shows the horizontal cross-section of the eye highlighting regions such as cornea, iris, pupil, lens, retina, fovea etc. Source [14].

Figure 2.2 shows the distribution of cones as a function of angle relative to the center of gaze. The density of cones, and hence the visual acuity, is very high in the fovea (0° x-axis in figure 2.2) and falls off rapidly as the angle increases. The fovea itself has a diameter of ≈1.5 mm and subtends an angle of approximately 2° of the visual field. This means that at any instant, less than 0.05% of our field of view is seen in high resolution.

We overcome this limitation by quickly scanning about the scene. Research has also shown that the peripheral vision is faster to respond to change in visual stimuli than the foveal vision. This is due to the fact that the optic fibers that carry signals from the peripheral regions of the retina to the primary visual cortex for processing are fast conducting while the optic fibers that carry signals from the fovea which are slower [16].

When viewing a scene for the first time, the low acuity peripheral vision of the HVS
Figure 2.2: Distribution of cones in the retina. Cones are densely packed in the center of gaze (fovea) and the density of cones falls off rapidly as angle from the center of gaze increases. The distribution of cones directly affects visual acuity. Visual acuity is highest in the center of gaze and falls off rapidly as the angle from the center of gaze increases. Adapted from Livingstone [17]

locates area of interest. The slower, high acuity foveal vision is then directed to fixate on these regions. Making eye movements to fixate on multiple regions help gather the scene information in high detail. Analyzing the fixations and their pattern reveals visual attention during task performance.

Human eye movements are broadly categorized into fixations, saccades, and smooth pursuits [1]. Other eye movement events include micro-saccades, ocular drifts, and ocular micro-tremors.

- **Visual Fixations**: A period of time when the eye is relatively stationary, maintaining viewer’s gaze on a single location to gather useful visual information.

- **Saccade**: A fast ballistic eye movement which is followed by a fixation. Saccades are useful eye movements that redirect the viewer’s gaze to regions of interest in the scene.
• **Smooth Pursuit:** The ability of the eye to follow a moving object. Humans cannot initiate a smooth pursuit without a moving visual target [18].

These eye movements play a significant role in human visual attention research. Section 2.2 provides a brief description on human visual attention and describes several computational models that have been developed for visual attention study.

### 2.2 Visual Attention

Vision plays an essential role to help humans perform a wide variety of tasks such as navigation, reading, driving, and interacting with objects. Attention in humans can be classified as *covert* and *overt*. Overt attention is the process of directing the fovea towards a desired target or stimuli to fixate on the object and gather information. Covert attention on the other hand is while focusing on an object gathering information on surrounding objects simultaneously (peripheral information). An example of covert attention is driving, where a driver while focusing on the road simultaneously covertly keeps tracks of his gauges, road signs and traffic lights. The theory behind covert attention is to quickly gather information on other interesting objects or features in the scene other than the one currently fixated. The reason for covert attention is due to the physiology of the eye that maps these saccades to other locations in the scene for information [19].

Visual attention can also be broadly classified as bottom up or top down. When humans look at a scene, they perceive some objects to be more interesting than others. There are certain objects in the scene that grab viewer attention over others. The drawing of visual attention in this fashion is termed as *bottom up* or *saliency-based* visual attention.
Focused attention can be thought of as a rapidly shifting spotlight and the areas focused are the salient regions in the scene. These salient regions can be represented as a 2-dimensional saliency map, which captures regions of high visual attention. However, human visual attention is not plainly a feed-forward spatially selective filtering process. There is also cognitive feedback to the visual system to focus attention in a *top-down* manner.

### 2.2.1 Bottom-up Saliency-Based Visual Attention

An early model on saliency-based visual attention was proposed in Treisman and Gelade’s *Feature Integration Theory* [20]. It was stated that combinations of visual features are important in directing attention during conjunction search tasks. Saliency maps, proposed by Koch and Ullman [21], are 2-dimensional representation of the most prominent locations in the scene. There were several visual saliency models proposed with digital images [22–24]. Itti et al. [25] provided a complete implementation and verification of the model proposed by Koch and Ullman and applied it to synthetic and natural scenes. Many algorithms have been proposed to formulate a model to compute visual attention on a given scene. Important and informative regions, surprising or out-of-place objects in a scene, or regions that reward most for a given task are some of the reasons that attract visual attention. These models have been extensively used for applications such as image-segmentation [26], scene classification [27, 28], object recognition [29, 30], and visual tracking [31, 32].
2.2.2 Top-Down Cognition-Based Visual Attention

Bottom-up cues are mainly based on characteristics of a visual scene. It is primarily stimuli driven and depend on the visual features of the scene [33]. Top-down models on the other hand are goal-driven or task-driven. Top-down visual attention models are determined by cognitive factors such as knowledge of the scene, expectations, rewards, tasks at hand, and goals. Bottom-up attention being feed-forward tends to be both involuntary and fast. On the other hand, top-down attention is slow, task driven and voluntary. Top-down attention is also refereed to as a closed-loop [34, 35]. The best known example of top down attention guidance is from Yarbus in 1967 [5, 36]. Yarbus showed that eye movement not only depend on the scene presented but also on current task on hand. Subjects were asked to watch a picture (a room with a family and an unexpected visitor entering the room) under different task conditions such as "what are the ages of the people?" or "estimate material circumstances of the family" or "free examination or free viewing." Figure 2.3 shows the scan path of a subject for various task while viewing the same stimulus image. Scene context and task at hand has shown to influence eye movements and visual attention.

2.2.3 Computational Visual Attention Models

Computational models have been developed to understand the influence of saliency and cognitive attention on visual search, object detection, scene classification, scene completion and task performance. Figure 2.4 shows the taxonomy of visual attention studies reproduced from Borji et.al [37]. Attention models have been classified as cognitive, Bayesian, decision theoretic, information theoretic, graphical, spectral analysis, and pattern classification etc.
Chapter 2 - General Background

Figure 2.3: Figure shows the experiment conducted by Yarbus in 1967 [5]. Image on top-left shows the picture of “Family and an unexpected visitor” and the scanpaths of a subject for each task in the experiment while viewing the image.

- **Cognitive models**: Models that closely relate the psychological and neurophysiological attributes of the human visual system to compute saliency. These models account for contrast sensitivity functions, perceptual decomposition, visual masking, and center-surround interactions [3, 38].

- **Bayesian models**: Models using Baye’s rule to detect object of interest or saliency regions by probabilistically combining extracted scene features with known prior knowledge of the scene or scene context [39, 40].
Figure 2.4: Taxonomy of visual attention studies. Boxes in blue indicate detailed focus in this project. Reproduced from [37]

- **Decision theoretic models**: Visual attention is believed to produce decisions on the state of the scene being viewed such that there is an optimal decision based on minimizing the probability of error. Hence salient features can be defined as best recognized classes over all other visual classes available [41, 42].

- **Information theoretic models**: Information theoretic models define saliency to be regions computed that maximizes information sampled from a given scene. The most informative regions are selected from all possible regions available in the scene [43, 44].

- **Graphical models**: Visual attention models are computed based on eye movements. These eye movements when treated as a time series and the hidden variables influencing these eye movements can be modeled using Hidden Markov Models,
Dynamic Bayesian Networks and Conditional Random Fields [45, 46].

- **Spectral analysis models**: A digitized scene viewed in the spatial domain can be converted to frequency domain and saliency models are derived based on the premise that similar regions in frequency domain imply redundancy. Such models are simpler to explain and compute by does not necessarily explain psychological and neurophysiological attributes of HVS [47–49].

These computational visual attention models use human eye movements as ground truth data for evaluation. Eye trackers provide an easy mechanism to gather human eye movements. Eye tracking technology plays a vital role in study of eye movements and enable researchers to understand and model human visual attention. Section 2.3 provides a brief description on history of eye tracking technologies and describe the working of video based eye trackers.

### 2.3 Eye Tracking

Eye tracking provides a mechanism for monitoring where our high acuity vision gets focused in a scene or on a display. Eye tracking systems first emerged in the late 1800s (see review of the history of eye tracking by Jacob and Karn [50]) and over the years various approaches have been used to track viewer gaze including magnetic coils embedded into contact lenses and electro-oculograms (EOG) which attempt to determine eye-position by taking advantage of the voltage difference between the retina and cornea. Most modern eye tracking systems are video-based - a video feed of the subject’s eye
is analyzed to determine the center of the pupil relative to some stable feature in the
video (such as the corneal reflection as shown in Fig 2.5). A brief calibration procedure
is usually necessary to establish the mapping of the subject’s eye position to locations
within the scene. The accuracy of video-based eye-trackers has improved in recent years
with many commercial systems reporting gaze position accuracy $< 0.5^\circ$.

Figure 2.6 shows the two commonly used eye tracking systems. The image on left shows
a video-based remote eye tracker and the image on right shows a head mounted eye
tracking system. The remote eye tracker systems presents the visual stimuli on a digital
display (label A). An infrared (IR)-based video eye tracker system fixed below the display
tracks the viewers eye movements (label B). An optional chin rest (label C) is shown in
the image to minimize head movements for better gaze accuracy. In the image on right,
the user wears a head mounted eye tracking system that allow for unrestricted head
movement. These mobile eye trackers are equipped with a front facing camera (label D)
that captures the scene the viewer is looking at, and a rear facing IR-based eye tracking
Eye tracking systems are primarily used to collect eye movement data during psychophysical experiments. This data is typically analyzed after the completion of the experiments. However during the 1980s, the benefits of real-time analysis of eye movement data were realized as eye-trackers evolved as a channel for human-computer interaction [51, 52]. Wearable eye-trackers began to emerge during the late 1990s [53, 54]. Today most of the commercial eye tracking companies offer head-mounted eye tracking solutions. Head-mounted eye-trackers allow researchers to capture information about the visual behavior and perceptual strategies of people who are engaged in tasks outside of the laboratory. They have already been used in a wide range of settings including driving [55], sports [56], geology [57], and mental health monitoring [58].
Chapter 3

Gaze Guidance for Problem Solving and Spatial Learning

When viewing traditional static images the viewer’s gaze pattern is guided by a variety of influences (bottom-up and top-down). The ability to direct a viewer’s attention has important applications in computer graphics, data visualization, image analysis, and training. Existing computer-based gaze manipulation techniques, which direct a viewer’s attention about a display, have been shown to be effective for spatial learning, search task completion, and training applications.

Jonides [59] explored the differences between voluntary and involuntary attention shifts and referred to cues which trigger involuntary eye-movements as pull cues. Computer-based techniques for providing these pull cues are often overt. These include simulating the depth-of-field effect from traditional photography to bring different areas of an image in or out of focus or directly marking up on the image to highlight areas of interest [60–64]. It is also possible to direct attention in a subtle manner. Studies on different cuing techniques have found a similar effect on target detection either using subtle or explicit
For example, the Subtle Gaze Direction (SGD) technique [67] works by briefly introducing motion cues (image space modulations) to the peripheral regions of the field of view. Since the HVS is highly sensitive to motion, these brief modulations provide excellent pull cues. To achieve subtlety, modulations are only presented to the peripheral regions of the field of view (determined by real-time eye tracking) and are terminated before the viewer can scrutinize them with high acuity foveal vision.

The work in this dissertation extends the Subtle Gaze Direction technique which is illustrated in Figure 3.1. Suppose that the goal is to direct the viewer’s gaze to some predetermined area of interest A. Let F be the position of the last recorded fixation, let \( \vec{v} \) be the velocity of the current saccade, let \( \vec{w} \) be the vector from F to A, and let \( \theta \) be the angle between \( \vec{v} \) and \( \vec{w} \). Either luminance modulation or warm-cool modulation is performed over the pixels in region A. Once the modulation commences, saccadic velocity is monitored using feedback from a real-time eye tracking device and the angle \( \theta \) is continually updated\(^1\) using the geometric interpretation of the dot product:

\[^1\text{SGD software has an eye-tracker gaze rate of 20Hz. This relatively low sampling rate works well in practice because research has shown that only 3-4 fixations occur per second [68]. A velocity threshold of 100 deg/sec was used to discriminate between saccades and fixations [69].}\]
\[ \theta = \arccos \left( \frac{\vec{v} \cdot \vec{w}}{|\vec{v}| |\vec{w}|} \right) \]  

(3.1)

A small value of \( \theta \) (\( \leq 10^\circ \)) or the current gaze location (\( \leq 5^\circ \) visual angle) indicates that the center of gaze is moving toward the modulated region. In such cases, modulation is terminated immediately. It is important to note that the modulation is terminated during the saccade to take advantage of the gap in our perception caused by saccadic masking (see Figure 3.1 inset). This contributes to the overall subtlety of the technique. By repeating this process for other predetermined areas of interest, the viewer’s gaze is directed about the scene. The modulations (operating frequency of 10Hz) are simply alternating interpolations of the pixels in \( A \) with black and white, in the case of luminance modulation, or with a warm and a cool color, in the case of warm-cool modulation.

Actively guiding viewers’ attention to relevant information has been shown to improve problem solving [61, 64, 70]. It has also been shown to enhance spatial learning by improving the recollection of location, size and shape of objects in images [62, 63, 71, 72]. Gaze manipulation strategies have also been used for improving performance on visual search tasks by either guiding attention to previously unattended regions [73] or guiding attention directly to the relevant regions in a scene [74, 75].

Gaze manipulation has high applicability in instructional environments. Numerous studies have been conducted to understand experts’ eye movements for specific tasks and to use their fixation sequence to guide novices during training. This has been done in tasks such as aircraft inspection [76] and optic disc examinations [77]. An interesting application of this approach is in the area of medical image analysis to understand the cognitive strategies employed by experts [78, 79]. Litchfield et al. [80] [81] showed that viewing an expert’s eye movements can help to improve identification of pulmonary nodules in
chest x-rays. Gaze manipulation has also been used to subtly guide novices along the scanpath of an expert radiologist as they try to identify abnormalities in mammograms [82]. Results of that study reveal that novices who were gaze guided performed significantly better than the control group (no gaze manipulation). They also continued to perform better once the training was complete and gaze manipulation was disabled.

The remainder of this chapter is organized as follows: the adaptive Subtle Gaze Direction is presented in section 3.1, impact of subtle gaze guidance on spatial information recall is discussed in section 3.2, the application of SGD in sequencing task on narrative art is presented in section 3.3, the experiment on applying subtle gaze manipulation for improved mammography training is described in section 3.4, a system to aid password recollection using subtle gaze guidance is discussed in section 3.5, and finally the approach for extending SGD beyond digital imagery into controlled real-world environments, is presented in section 3.6.

### 3.1 Adaptive Subtle Gaze Direction Technique

Adaptive subtle gaze direction technique is based on the original SGD approach developed by Bailey et.al [67]. Several changes were made to ensure that the technique is both fast and accurate for real-world training and learning applications. Adaptive SGD uses a 2D Gaussian pulse for modulation as shown in the equation below.

\[
f(x) = \frac{K}{2\pi \sigma^2} \exp\left(\frac{(x + y)^2}{2\sigma^2}\right)
\]

Figure 3.2 shows the Gaussian functions for the adaptive subtle gaze modulation. Images on top in Figure 3.2 show modulations with different spread on the observer’s field of view.
Figure 3.2: Figure shows the Gaussian modulations used in the adaptive subtle gaze approach. Images on top show the 2D Gaussian plot with different SD (σ) values for varying the spread of the pulse. Images at the bottom show the positive and negative cycle happening at different time period to ensure randomness in the modulation.

view. The spread of the Gaussian pulse is increased over time until the viewer makes a saccade towards the desired target location. This functionality is achieved by choosing the Gaussian function with different standard deviations (σ). The modulation intensity is also gradually increased by changing the $k$ value in the equation 3.2. This ensures that the modulation attracts the visual attention of the viewer if the target location happens to be in the extreme peripheral location in the observer’s field of view. Finally, the modulations are performed at different time periods instead of a constant time period chosen in the original SGD technique. This ensures randomness instead of a constantly alternating positive and negative pulse.
3.1.1 Experiment Design

The modulations presented to subjects vary in spread, amplitude and time. The modulations were performed on the intensity channel of the image. This ensured that the modulations were equal when applied on gray-scale or RGB images. The RGB images were converted to HSV color space to facilitate modulations on the intensity channel. The adaptive SGD modulation was programmed as a separate thread to reduce latency and later upgraded to run on the GPU to ensure a modulation frequency of 60Hz. This was essential to match the operating frequency of the display system. Thus, the adaptive SGD approach ensures that the viewer makes a saccade towards all target locations and the modulations are both subtle and fast.

To evaluate the effectiveness of adaptive subtle gaze guidance technique, a pilot study was conducted with six participants (2 females, 4 males). Participants who volunteered for the study were between the ages of 18 and 23. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. Participants were randomly assigned to one of two groups:

- **Static group:** Participants (three) were presented with a randomized sequence of the 10 images with no modulation. This group served as the control group for the experiment.

- **Modulated group:** Participants (three) were presented with a randomized sequence of the 10 images with modulations at pre-selected image locations. These locations, manually selected by the researchers, included areas that are not visually significant - low contrast, low detail, low color saturation, and uninteresting objects. These are areas that an observer would not ordinarily attend to.
3.1.2 Results and Discussion

To evaluate the effectiveness of adaptive SGD technique gaze data, activation-time, target acquisition and gaze pattern change was analyzed between the control-group and the gaze guided group. The adaptive SGD approach was effective in the activation time and target acquisition. There was no significant change in the overall gaze pattern. However the frequency of modulation was significantly improved from 10Hz to 60Hz. Figure 3.3 shows the gaze distribution results obtained using the original Subtle Gaze Direction approach (The adaptive SGD approach produces similar results). The heat map on the left indicates the natural viewing behavior of the participants from the control group. The heat map on right indicates the gaze behavior of the participants
from the modulated group. The white crosses in the heat map indicate the modulated regions and it is evident that the viewing pattern of the participants has been changed by directing their gaze in a subtle manner.

3.1.3 Conclusion

The adaptive SGD technique improves on the original SGD approach by gradually increasing the spread and intensity of modulations until the viewer makes a saccade towards the target location. The modulation frequency was also increased to 60Hz to match the operating frequency of the display by processing the stimuli image using multi-threaded GPU-based approaches. This ensures more subtlety and provided opportunities to evaluate the applications of adaptive SGD in aiding short-term memory, information recall, training and learning, extending gaze guidance to real-world and 3D virtual environments. In further sections, the term SGD is used instead of adaptive SGD for ease, and all references to modulations use the adaptive approach for subtle gaze guidance.

3.2 Impact of Subtle Gaze Direction on Short-Term Spatial Information Recall

Everyday tasks such as retrieving items from specific locations or navigating a familiar route rely on accurate spatial understanding of the environment. Spatial memory enables a variety of species, including humans, to maintain stored information about the location of objects in their environment [63, 83, 84]. Spatial memories of the location, shape, and number of items in a scene become reinforced and more accurate in direct proportion to the amount of time spent in an environment, or the frequency with which tasks are
completed. Spatial memory becomes critical for tasks that occur in the absence of clear visual information such as navigating through a building in the dark. Humans also rely on spatial memory for simpler tasks such as interacting with images or videos viewed on 2D displays. Over time, for example, people learn menu options or icon locations, the path an object takes through a video sequence, or how to navigate through various levels of video games. This type of spatial (also spatial-temporal for video) learning contributes to the overall understanding of a scene.

A key characteristic of accurate scene interpretation is the amount of information that can be extracted from a scene and retained in memory. Visual memory can be categorized into Visual Short-Term Memory (VSTM), Visual Long-Term Memory (VLTM), and Iconic (sensory) Memory [85]. VSTM is limited in terms of storage capacity but creates representations very rapidly, which can then be used to inform ongoing cognitive tasks. On the other hand, VLTM boasts virtually unlimited storage capacity and over time forms detailed representations. For this project we are interested in VSTM, in particular whether/how we can use gaze manipulation to help prioritize the information it acquires in order to better perform a spatial recall task. It is well known that the content of VSTM is highly dependent on where the viewer attends in the scene [86].

3.2.1 Experiment Design

An experiment was conducted to investigate the impact of SGD on short-term spatial information recall. Participants viewed a randomized sequence of images. Following each image, they were presented with a blank screen and asked to recall the location of specific objects or regions. They were instructed to use the mouse to draw the smallest rectangles that bounded each target object or region. Their input was later analyzed to
determine how accurate their short-term spatial recollection was in terms of number of targets, location, and shape.

3.2.1.1 Stimuli

Stimuli were presented on a 22 inch widescreen monitor, operating at 60 Hz with a resolution of 1680 x 1050. The stimuli consisted of 28 images (3 training images and 25 test images) compiled from various sources. The images ranged from simple scenes with a few objects to complex scenes with many objects. The number of objects or regions that the participants were asked to recall for each image ranged from 1 to 9 with each number in the range represented. We used Miller’s observation, [87] which states that the average human can only hold $7 \pm 2$ items in working memory to establish the upper limit of 9 for the experiment.

3.2.1.2 Participants

30 participants (4 females, 26 males), between the ages of 18 and 35 volunteered to participate in this study. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. Participants were randomly assigned to one of two groups:

- **Static group:** 10 participants were presented with a randomized sequence of the 25 test images without the use of gaze direction. This group served as the control group for the experiment.
• **Gaze-directed group:** 20 participants were presented with a randomized sequence of the 25 test images with gaze-direction turned on. For each image presented to the participants in this group, gaze-direction was randomly selected to either:

  – Direct the viewer’s gaze to randomly selected regions of the image away from the correct targets.

  or

  – Direct the viewer’s gaze to the correct target regions of the image.

Counterbalancing was used to ensure that every image appeared equally often in both gaze-directed conditions (i.e. modulation at correct targets and modulation away from correct targets).

This approach facilitates the following comparisons for each image:

• 10 viewers from the static group versus 10 viewers from the gaze-directed group who were correctly guided.

  and

• 10 viewers from the static group versus 10 viewers from the gaze-directed group who were incorrectly guided. i.e. gaze was directed away from the information most pertinent to accurately responding to the subsequent recall task.

We hypothesized that using SGD to guide the viewer’s focus to the correct target regions of the image would improve short-term spatial recall accuracy compared to static viewing. Similarly, we hypothesized that using SGD to guide the viewer’s focus to incorrect target regions would lower short-term spatial recall accuracy compared to static viewing.
3.2.1.3 Procedure

Since we are interested in exploring the impact of SGD on VSTM, each image was shown for only 5 seconds and then replaced with a blank screen. While the blank screen was being displayed, an audio question about the spatial content of the image was played. The questions were recorded in a normal voice by a male native English speaker. Sampling rate for the audio questions was 44.1 kHz. This approach is preferred to displaying the question as text on-screen as this may disrupt the participant’s short-term memory of the image [88].

The participants responded to the questions by using the mouse to draw rectangular regions on the blank screen. Three images from the complete set were used in a brief training session to ensure that the participants understood the procedure for completing the experiment. During the training session, participants were encouraged to ask questions and were able to view their solution and the correct solution after each image (see Figure 3.4).

Subtle gaze guidance was implemented as described in 3.1. The eye-tracker used in this study is a SensoMotoric Instruments iView X Remote Eye Tracking Device operating at 60 Hz with gaze position accuracy < 0.5°. A chin rest was not used in this eye tracking study. Assuming that there were \( n \) correct solutions and \( m \) participant responses for a given image, we define the following measures for spatial recall accuracy:

- **Counting error** is the difference between the number of correct regions in an image and the number of regions submitted by the participants. There is no penalty for incorrect location or shape of the rectangular regions that the participants
Figure 3.4: One participant’s solution (red) and correct solution (blue) for an image from the training set. Participants were shown the image for 5 seconds then presented with a blank screen and asked to recall the location of each duck in the image by drawing rectangular regions on the blank screen.

submit. Counting error is defined as follows:

$$|n - m|$$ \hspace{1cm} (3.3)

Where \( n \) is the number of correct regions in an image and \( m \) is the number of regions submitted by the participants.

- **Location error** is a measure of how close the participant’s responses were to the actual targets. There is no penalty for incorrect count or shape of the rectangular regions that the participants submit. Location error is defined as follows:

$$\sum_{i}^{\min(n, m)} \left( \sqrt{(x_{ai} - x_{bi})^2 + (y_{ai} - y_{bi})^2} \right)$$ \hspace{1cm} (3.4)

where \( i \) is the smaller of \( n \) and \( m \) and \((x_{ai}, y_{ai})\) and \((x_{bi}, y_{bi})\) represent the centroids of the \( i \)th closest pair of rectangles chosen from the set of actual solutions
and participant solutions.

- **Shape error** is a measure of how different the widths and heights are of the actual solutions and the rectangular regions that the participants submit. There is no penalty for incorrect count or location. Shape error is defined as follows:

$$\sum_{i}^{\text{min}(n, m)} (|\text{width}_{ai} - \text{width}_{bi}| + |\text{height}_{ai} - \text{height}_{bi}|)$$  \hspace{1cm} (3.5)

where $i$ is the smaller of $n$ and $m$ and $\text{width}_{ai}$ and $\text{width}_{bi}$ and $\text{height}_{ai}$ and $\text{height}_{bi}$ are the widths and heights of the $i$th closest pair of rectangles chosen from the set of actual solutions and participant solutions.

Location error and shape error are undefined in cases where the participants did not submit any solutions. This occurred for 15 of the 250 data points for the static condition, 13 or the 250 data points for the SGD at incorrect locations condition and 6 of the 250 data points for the SGD at correct locations condition. These data points were not included in the computation of location error and shape error.

### 3.2.2 Results and Discussion

We measured the impact of SGD on short-term spatial information recall by computing the participants’ counting error, location error, and shape error. In summary, we observed the following effects:

- SGD to correct targets results in a significantly lower counting error compared to static viewing.
Figure 3.5: Average counting error for participants from the static group and gaze directed group. Error bars represent standard error.

- SGD to correct targets results in a significantly lower location error compared to static viewing
- SGD to incorrect targets results in a significantly higher location error compared to static viewing
- SGD has no significant impact on shape error

3.2.2.1 Counting Error

Figure 3.5 summarizes the average counting error for the participants from each group. Counting error for the static group averaged 1.488 targets while the averages for the gaze-directed group were 1.448 targets (modulations at incorrect locations) and 0.76 targets (modulations at correct locations). These values were obtained by averaging the counting error for all participants in a group over all images.

The differences in the averages show that SGD to correct locations results in a lower counting error compared to static viewing. An independent-samples t-test revealed that
Figure 3.6: Average counting error for participants from the static group and gaze directed group as number of target regions increase.

this effect was significant and not due to chance:

\[ t(498) = 4.640; p < 0.05 \]

Figure 3.6 shows how the average counting error varies as the number of regions the participants were asked to recall increases. As expected, the counting error increases as the recall task becomes more difficult. Note, however, that the counting error for SGD to correct targets is consistently lower than that of static viewing and that of SGD to incorrect targets.

### 3.2.2.2 Location Error

Figure 3.7 summarizes the average location error for the participants from each group. Location error for the static group averaged 134 pixels while the averages for the gaze directed group were 161 pixels (modulations at incorrect locations) and 99 pixels (modulations at correct locations). These values were obtained by averaging the location error for all participants in a group over all images.
Figure 3.7: Average location error in pixels for participants from the static group and gaze directed group. Error bars represent standard error.

The differences in the averages show that SGD to correct locations results in a lower location error compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

\[ t(477) = 4.123; p < 0.05 \]

The differences in the averages also show that SGD to incorrect locations results in a higher location error compared to static viewing. An independent-samples t-test revealed that this effect was also significant and not due to chance:

\[ t(470) = -2.526; p < 0.05 \]

Figure 3.8 shows how the average location error varies as the number of regions the participants were asked to recall increases. Ignoring the outlier (actual number of targets = 1), the plot shows that location error remains fairly constant as the number of targets increase with the location error for SGD to correct locations generally being the lowest and location error for SGD to incorrect locations generally being the highest.
Figure 3.8: Average location error for participants from the static group and gaze directed group as number of target regions increase.

Figure 3.9: An image from the test set that was potentially ambiguous. Several participants confused squashes and peppers possibly due to lack of familiarity. This resulted in a spike in location error for this image. Image courtesy of Natalie Maynor [89].

Closer examination of the outlier revealed that the irregularity was due to a potentially ambiguous trial in our test set. Figure 3.9 shows the image in question. Participants were asked to recall the location of each pile of yellow peppers. We used the word “each” for every question in the experiment to avoid implying the correct count. We observed that several participants selected the yellow squashes instead of the yellow peppers possibly
due to lack of familiarity. This led to a spike in the location error associated with this image. It is interesting to note however, that SGD to the correct location seems to have helped to resolve this ambiguity for the gaze-directed group.

### 3.2.2.3 Shape Error

Figure 3.10 summarizes the average shape error for the participants from each group. Shape error for the static group averaged 132 pixels while the averages for the gaze-directed group were 142 pixels (modulations at incorrect locations) and 120 pixels (modulations at correct locations). These values were obtained by averaging the shape error for all participants in a group over all images.

The differences in the averages suggest that SGD to correct locations results in a lower shape error compared to static viewing. However, an independent-samples t-test revealed that this effect was not significant:

\[ t(477) = 1.310; p > 0.05 \]

In this experiment, SGD was used to direct gaze to locations that were clearly within the boundaries of the target regions or objects. We speculate that using SGD to also guide viewer gaze along the bounding box of the target may further improve short-term shape recall. No clear trend was observed in the average shape error as the number of target regions increased.

### 3.2.2.4 Percentage Gaze Time

We have observed that SGD to correct image locations significantly improves accuracy of target count and spatial location recall. To gain a better understanding of why these
Chapter 3 - Gaze Guidance for Problem Solving and Spatial Learning

![Average shape error for different groups of participants](image1)

**Figure 3.10:** Average shape error in pixels for participants from the static group and gaze directed group. Error bars represent standard error.

![% gaze in target regions for different groups of participants](image2)

**Figure 3.11:** Percentage gaze time spent within the target regions for participants from the static group and gaze-directed group. Error bars represent standard error.

effects occur we examine how the percentage gaze time spent within the target regions is affected by SGD. Figure 3.11 shows the percentage of total gaze time spent within the target regions for the different groups of participants. For the static group, 8.7% of gaze time was spent within the target regions. For the gaze-directed group, 12.5% (which represents a 43.7% increase) and 15.2% (which represents a 74.7% increase) of the gaze time was spent within the target regions for modulations at incorrect locations and modulations at correct locations respectively.
The differences in the averages show that SGD to correct locations results in more gaze time spent within the target regions compared to static viewing. This observation was expected and an independent-samples t-test confirms that the difference in gaze time in the target regions was significant and not due to chance:

\[ t(498) = -4.135; p < 0.05 \]

Interestingly however, the differences in the averages also show that SGD to incorrect locations results in more gaze time spent within the target regions compared to static viewing. An independent-samples t-test also confirms that the difference in gaze time in the target regions was significant and not due to chance:

\[ t(498) = -2.605; p < 0.05 \]

This observation was somewhat unexpected. We speculate that the increase in gaze time in the target regions is due to the fact that SGD to various incorrect locations helps to distribute the viewer’s gaze more evenly across the image. This causes more of the viewer’s scan-paths to intersect with the target regions. A similar observation was also made in another study involving SGD [74]. The researchers noted that the presence of “distractors” (i.e. modulations at incorrect regions) helped to spread the viewer’s gaze across images and led to improved performance on a search task compared to static viewing. We plan to explore this further in future experiments.

For this study, the fact that both correct and incorrect modulations resulted in increased gaze time in the target regions but only correct modulations led to improved accuracy of target count and spatial location recall seems to suggest that the modulations themselves serve as triggers for retaining information in our short-term memory.
3.2.3 Conclusion

In this study, participants were asked to recall the location of objects in images. We investigated if using SGD to guide attention to these regions would improve spatial information recall. Results showed that the influence of SGD significantly improved accuracy of target count and spatial location recall. The effect was observed on a wide variety of images ranging from simple scenes with a few target regions to complex scenes with many target regions. The results from this experiment motivated the work to study the impact of SGD on sequencing tasks.

3.3 Directing Gaze in Narrative Art

Narrative art tells a story, either as a moment in an ongoing story or as a sequence of events unfolding over time. A synoptic narrative depicts a single scene in which a character, or characters, are portrayed multiple times within a frame to convey that multiple actions are taking place. This can cause the sequence of events to be unclear within the narrative. Synoptic narratives typically provide visual cues that convey the sequence, but it still might be difficult to decipher for those unfamiliar with the story. The process is best illustrated with an example. Figure 3.12 (top) shows a synoptic painting titled “The Tribute Money” by Renaissance artist Masaccio. This painting describes a story from the Gospel of Matthew, in which Jesus directs Peter to go to the river and retrieve a coin from the mouth of a fish in order to pay the temple tax. The optimal way to visually navigate this piece is to begin in the center with the tax collector demanding the money. Jesus, surrounded by his disciples, instructs Peter to retrieve the money from the mouth of a fish. By moving their gaze to the left of the painting (perhaps counter-intuitive to western civilization who normally read left to
Figure 3.12: Above: “The Tribute Money”, by Masaccio tells the story of Jesus and the tax collector. The piece should be viewed in the following order: center, left, then right. Current web-browser based educational tools use text pop-ups with interruptive rectangular outlines to highlight important information in a visual narrative. This not only distracts the viewer from appreciating the image, but also breaks up the image into smaller pieces so it is not viewed in a holistic manner. The red colored rectangle destroys the visual experience by superimposing a distracting overlay on the original painting.

(right) viewers notice Peter executing Jesus’ instruction. The viewer’s eyes next need to travel to the extreme right of the painting to view the third episode in which Peter pays the tax collector. At the time it was painted, audiences were conditioned to recognize repeated elements in a frame and identify panels, thereby intuitively understanding the intended order in which each episode of the painting was to be viewed. However, our ability, as artists and audiences, to correctly “read” these paintings may not be so accurate in present day because our visual literacy is not conditioned to follow the viewing pattern the artist intended. In the 15th century the audience would understand
that there are multiple episodes in this painting, and also in which order to view these
panels in order to comprehend the story. Web-based solutions do exist which manipulate
a digital representation of a painting using strong outlines, or interruptive text over the
image to explain where the viewer should direct their gaze (see Figure 3.12, [90]). While
these represent a promising initial approach, a more elegant solution would not disrupt
interrupt the visual experience of the audience.

The marriage of technology and art appreciation is not new. Several existing applications
have successfully been applied in the Art domain [91] [92] [93] [94] [95] [96] [97]. To date
however, few have proposed eye tracking as an added dimension. The novelty of this
approach lies in the eye tracking and in attracting and directing the gaze to the correct
region of the artwork in a sequence that will encourage appropriate visual navigation
and understanding of the image and strengthen observation skills.

Employing gaze direction techniques would allow the viewer to see the actual painting
with areas of interest accentuated in a manner which preserves the visual experience by
acknowledging the artist’s intent. In this section, we investigate the use of SGD as an
aid to navigate narrative art. This goal of this work is to satisfy the need to display
information in a manner that minimizes disruption to the viewer, but can accurately
direct gaze to certain locations of an image, in a specific sequence. In other words, how
well can SGD direct viewers’ gaze to multiple image locations in a specific sequence.

3.3.1 Experiment Design

The aim of the experiment in this section is to determine to what extent SGD lends
itself to aid observers in extracting the intended sequence of events from regions of an
episodic image. Participants viewed a sequence of images, each of which contained three
or more panels intended as episodes. The intended viewing order of these panels is not always immediately clear. Art history research provides the narrative for each art piece, from which panels are determined [98]. Panels (in each image) were manually selected as rectangular regions which enclosed the relevant portion of the image that conveyed an episode of the story. In the non-control group, after viewing the image for a short period of time, relevant panels were highlighted using SGD at the panel center. Participants then indicated the order they perceived to be the correct viewing order by clicking on image sections outlined with boxes. We compare performance using SGD with performance under normal viewing conditions.

### 3.3.2 Stimuli

Eleven images served as stimuli for the experiment, two of which were used for observer training. In each image, episodes or panels were identified and served as targets for SGD. The number of episodes (panels) varied from painting to painting, ranging from three to seven. The size of panels also varied within each image (see Figure 3.13), initially we were concerned that fixations may be artificially increased in proportion to panel size, but this did not turn out to be the case.

Presentation order was randomized to minimize the introduction of learning effects. Images were presented for a period of time proportional to the number of episodes in the image and were displayed on a 22 inch widescreen monitor, operating at 60 Hz with a resolution of 1680 x 1050. Stimuli transitioned directly from one to the next, however, no action was taken until a finite amount of time had passed. Image sizes varied in the experiment, in cases where images size was smaller than the viewing screen, width and height were maximized to fit screen resolution and a black border was added.
Participants were seated in front of a computer screen in a well-lit room. Using a SensoMotoric Instruments iView X Remote Eye Tracking Device operating at 250 Hz with gaze position accuracy \(< 0.5^\circ\), data pertaining to fixation position and saccades were recorded for each participant. A chin rest was not used in this eye tracking study. After a brief calibration phase, each observer underwent a short tutorial session to familiarize them with the experimental procedure and user interface. Questions were encouraged during the tutorial session but no data was collected. Participants were then presented with each of the nine art works in a random order. Image complexity varied, as did the number of panels. The two groups are as follows:

- **Group 1: Normal Viewing Conditions**: No actions were applied to the images, in other words images were viewed normally with no modulations. This group served as the control group for the experiment.

- **Group 2: Subtle Modulation**: SGD was employed to highlight the target panel regions in the intended viewing order in an effort to aid in visual navigation. Gaze manipulation was implemented as described in 3.1. The modulations were placed in the center of the panels.

Thirty-six participants were assigned randomly to one of the two groups. Participants were volunteers from a group of undergraduates. All had normal, or corrected-to-normal, vision and were naive to the purpose of the experiment. Viewing time for each image varied in direct proportion to the number of panels present. Each image was presented for \(t\) seconds before the user was allowed to respond. For the control group, Group 1, \(t\) was chosen to be equal to the number of regions in the image. In Group 2, \(t\) is the time taken to guide the viewers exactly once through the correct sequence of regions. Previous studies [67], [99] revealed that SGD modulations typically attracted gaze within
Figure 3.13: An exemplar image showing all of the panels which contribute to the narrative. Observers first viewed the image without the panels highlighted. Once a certain amount of time had elapsed, participants then clicked on the panels in the order they believed matched the order of the story being told. The modulations for the SGD group (Group 2) were placed at the center of each panel.

0.5 seconds. To ensure that we had comparable viewing times between both groups a 0.5 second delay was added between successive modulations.

After $t$ seconds, the relevant regions were highlighted with rectangles and the mouse activated to allow the users to respond. Both groups of participants were instructed to click on the highlighted regions in the order they believed the story unfolds. Each participant reported an order which they believed matched the intended sequence of the story in the art work.
3.3.2.1 Analysis of Data

In addition to recording eye-movements for each participants, each participant reported an order for each image, based on their understanding of panel sequence within that image. We needed a robust mechanism to compare accuracy of performance between the two groups. Levenshtein distance \([100][101][102]\) is a string metric, developed in the field of information theory and computer science to compute differences between sequences. Levenshtein distance provides an appropriate measure to compare distances between ordered sequences, such as those recorded during our experiment. To accurately compare sequences using Levenshtein distance the correct (intended) viewing order of each image is converted into a string sequence. All responses from each participant are also converted to an appropriate string sequence in order to facilitate comparison to the correct sequence. Since the number of relevant regions varies across the images we normalize the distance measure computed for each image by dividing by the number of panels. The normalized Levenshtein distance \(L\) between the correct sequence \(S_{\text{correct}}\) and user sequence \(S_{\text{user}}\) is as follows:

\[
L = \frac{\text{Levenshtein Distance}(S_{\text{correct}}, S_{\text{user}})}{\# \text{ of Panels}} \times 100
\]  

(3.6)

For example, let the correct panel order be [ABCDE] and let [ACBDE] denote the participant’s response. The number of panels in this image is five. Using Equation 3.6, we obtain a normalized Levenshtein distance value of 40. A distance of 0 would indicate no difference, whereas a distance of 100 would indicate maximal distance.
3.3.3 Results and Discussion

The predicted sequence of panels reported by each participant for each image was recorded. Normalized distances for each image were compared to the actual intended sequence for that image using the distance metric expressed in Equation 3.6. The calculated normalized Levenshtein distance measures between conditions showed differences across the two groups with a mean distance measures of 57.32 and 34.79 for groups 1 and 2 respectively, as illustrated in Figure 3.14. These values were calculated by averaging the normalized Levenshtein distance \( L \) for all the participants in a group over all the images used in the experiment. This implies that participants from Group 2 (guided by SGD) consistently proved to be more accurate at predicting the intended sequence of panels contained in narrative art when compared to Group 1, the control group (static viewing). This measure indicates that, for example, when viewing a narrative art image containing ten panels, the static viewing group will incorrectly predict the order of approximately 5 – 6 panels, while the gaze directed group, Group 2, will return a prediction with only 3 – 4 panels out of the sequence. An independent-samples paired \( t \)-test suggests that this was a significant effect:

\[
t(316) = 1.9675; p < 0.05
\]  

Figure 3.15 shows the average L value for each image across all the participants in each group. This analysis gives some intuition on the influence of the number of panels over the accuracy in detecting the correct sequences in the narrative art. The images in the graph are arranged in the increasing order of number of panels. In all nine images, the average L value indicates that the participants in Group 2, the gaze directed group, predict panel order more accurately than the participants from Group 1, the static
Figure 3.14: Normalized Levenshtein distance measure between Group 1 (static viewing group) and Group 2 (gaze directed group). Error bars represent one standard error. The graph shows that Group 2 participants, who viewed SGD images were able to predict the intended viewing order of panels more accurately than those in Group 1 that did not have the benefit of SGD as a gaze direction aid.

Figure 3.15: Normalized Levenshtein distance measure between static viewing group and gaze directed group for each image. The x-axis indicates the number of panels in each image. The error bars represent one standard error.
Number of panels & Independent t-test (Group 1 V Group 2) \\
--- & --- \\
3 & $t(33) = 2.0322$ \\
4 & $t(33) = 2.0346$ \\
5 & $t(33) = 2.0345$ \\
6 & $t(33) = 2.0340$ \\
7 & $t(33) = 2.028$ \\
26 & $t(33) = 2.0364$

Table 3.1: Independent t-tests indicate significant differences in the ability to correctly predict intended panel sequences for images with vary in numbers of panels. In each case, $p < 0.05$.

viewing group. This also shows that the gaze directed group performed better than the static group for images having relevant regions varying from 3 to 26. In eight of the nine images used in the study, this result was shown to be significant. Independent-samples t-tests reveal that this effect was significant and not due to chance, the t-test results for images with relevant regions 3, 4, 5, 6, 7, and 26 are shown in Table 3.1.

The results did reveal a single anomaly where the t-test did not show a significant difference between groups. This anomalous image revealed no significant difference in performance between the two groups. Further inspection showed that in this image, the artist has gradually decreased the luminance of the narrative art over the story. This analysis was possible as the same characters appear over multiple regions in the image. We reason that this luminance change in itself would provide a strong enough visual cue to enable the participants in Group 1 to correctly navigate the story. This also suggest that luminance changes could serve to guide direct gaze in imagery. This phenomenon is a topic for future research.

To further illustrate the success of SGD we present a single example. Figure 3.16 shows the images for the static viewing and the gaze directed group respectively. Images are placed side-by-side for comparison.
Figure 3.16: This Figure shows the scan paths and heat-maps for one participant from each of the two groups (no modulation and SGD). Image A & C represent data collected from Group 1, while images B & D were collected from Group 2. Rectangular highlighted regions denote panels. Blue numbered circles indicate the correct viewing order of panels within the image. As can be seen from the images, gaze distribution and fixations are more accurately aligned with panel (modulated) regions in the SGD condition.

Image A depicting the scan path of the viewer’s gaze over the static images shows that the viewer’s gaze does not coincide with all of the relevant story panels. Contrast this with image B, which shows the scan path over the SGD enhanced image. This image reveals a more coherent scan path in terms of visitation to each relevant panel.

Comparing heat maps reveals a similar story. The heat maps represent the amount of time spent fixating in each image region. Figure 3.16, image C reveals that most fixations fall to the left of the image, and the distribution does not encompass the story panels. Conversely, examination of the heat map for image D (SGD) indicates that viewer fixations are distributed over the story panels.

Examining the L value measures (as described in Section 3.3.2.1) for this single image, the Group 1 (no SGD) participant’s value is 71.45 compared to 28.57 for the participant.
from Group 2 (the gaze directed group). Thus the heat map and scan path analysis not only reflect the increase in the gaze coverage and attention to all relevant regions of the image for the gaze directed group over the static viewing group, but also correspond well with the L-value metric chosen to compare performance.

It is important to note that Figure 3.16 serves as a representative example of a consistent trend across all nine images viewed. This analysis reveals that, without SGD, not only did participants fail to view all of the story panels, but they failed to fixate on all the relevant story panels. The exact opposite is true for those images presented with SGD applied to the story panels, giving a high level of confidence in the success of applying SGD to subtly reveal an intended viewing sequence.

In summary our results show that gains can be made in task performance when modulation is employed to direct gaze to target a sequence of panels in a specific order. Even if the modulations are noticed to some degree, there is still an increased accuracy in task performance. This seems to hold true over a range of images, and over a range of panel numbers. For applications that require a specific viewing order for understanding or performance, SGD can serve as a subtle aid to boost accuracy of performance on a sub-image ordering task.

### 3.3.4 Conclusion

We presented an experiment to compare task performance in digital images across two groups of stimuli. In one group no image alterations were used (Group 1), in the second group small modulations were applied to image panels in an effort to direct the viewers’ gaze (Group 2). The participant’s task was to specify the order of panels (contained in episodic art pieces) which revealed the intended story. The results indicate that
using a subtle gaze direction technique, which modulates the appropriate panel in the intended sequence, does indeed improve the precision of panel ordering. The performance difference between the two groups was shown to be significant.

We have shown that SGD can improve performance on a within-image panel ordering set without noticeably disrupting the visual experience of the image. This technique can be applied to help guide gaze in complex applications where viewing order is critical to understanding, such as in story telling, or training task performance where sequence of operations is important, for example, construction instructions. To test the application of one such technique, a study was designed to see if guiding novices along an expert’s scanpath will enable them to perform better than a control group.

3.4 Subtle Gaze Manipulation for Improved Mammography Training

The American Cancer Society estimates that 246,600 new cases of breast cancer were diagnosed in the US in 2015-2016 [103]. The statistics also show that breast cancer is the leading form of cancer among women. Early diagnosis and treatment of breast cancer are essential to maintaining the patients long-term health. In addition to breast self-exams and clinical breast exams, mammograms play an important role in the early detection of breast cancer. Radiologists undergo extensive training to become proficient at reading mammograms. In the US, an expert radiologist typically completes four years of undergraduate study, four years of medical school, one year of internship, four years of residency training and one or two years of additional fellowship training. Despite advances in technology, radiological training still uses the conventional approach of having a trainee work alongside an expert radiologist.
The study of expert gaze patterns within the medical field is not limited only to breast cancer research and mammograms. Krupinski [104] summarizes work in this area and discusses the importance of perception research in medical imaging. Carmody et al. compare fixations between instructors and radiology residents for lung scan images [105] and chest scans [106]. Sowden et al. [107] used eye tracking to study how perceptual learning affects a subject’s ability to detect features in general X-ray images. Litchfield et al. [80, 81] showed that viewing an expert’s eye movements can help to improve identification of pulmonary nodules in chest x-rays and aids problem solving.

Understanding the diagnostic process used by radiologists when searching for cancerous regions in mammograms is an interesting and active area of research. Studies have been conducted to analyze the fixation pattern of expert radiologists when viewing mammogram images. Kundel et al. [108] recorded and analyzed the fixations of expert radiologists from three independent institutions as they studied cancerous regions on mammograms. Their work provides useful information about the duration of expert viewing and identifies several regions of interest in digital mammograms. Mello-Thoms et al. [109] also analyzed dwell time and number of fixations by experts during scans for breast cancer in mammograms using both head mounted and remote eye tracking devices. Krupinski et al. [110] studied decision making among mammographers and radiology residents with gaze duration as the key parameter.

Currently most radiological training programs still use the conventional approach of having a trainee work alongside an expert radiologist. We propose a novel computer-based training technique that uses gaze manipulation to guide novices along the recorded scanpath of an expert radiologist. Computer-based training systems and workstations have become more popular in the medical field with the emergence of digital medical images and improved computer graphics and visualization techniques [111–117].
In this section we present a novel training technique that uses subtle gaze manipulation to guide novices along the recorded scanpath of an expert radiologist. We hypothesized that guiding a novice along the scanpath of an expert radiologist would improve the likelihood that the novice correctly identifies the abnormalities in the mammograms. To test this hypothesis, we designed an experiment to explore whether SGD is capable of improving the efficiency of digital mammography training.

### 3.4.1 Experiment Design

In this section we describes an experiment conducted to investigate if subtle gaze manipulation is capable of improving the efficiency of digital mammography training. During a training session, participants viewed a randomized sequence of mammogram images and were asked to identify what they considered to be irregularities. There were four groups of participants. One group viewed the images without the use of gaze manipulation. They served as the control group for the experiment. For the other groups, the SGD technique was used to guide the participants in various ways. Two follow-up sessions were completed without gaze manipulation to determine if the participants became sensitized to the training method used.

#### 3.4.1.1 Stimuli

The database of mammogram images used in this study was provided by the Mammographic Image Analysis Society (MIAS) [118]. The database contains pairs of mediolateral images from 161 patients along with a separate text file containing information collected from an expert radiologist such as the x,y image coordinates of the center
of abnormalities and the approximate radius in pixels of a circle enclosing the abnormality. What is missing however, is the experts scanpath when trying to locate these abnormalities. We hired our own expert radiologist to view a subset of 65 images from the database and to mark any abnormalities present by drawing a circle enclosing the abnormality. The experts scanpath was recorded during this process and later used to guide one group of the novice participants.

Stimuli for the experiment were presented on a 22 inch widescreen monitor, operating at 60 Hz with a resolution of 1680 x 1050. The stimuli consisted of the 65 pairs of mammogram images that were viewed by the expert. Five of these images were used for an initial tutorial session to familiarize the participants with the software interface. Twenty images were used for a training session, twenty images were used for a short-term follow-up session, and twenty images were used for a long-term follow-up session. For all sessions, the participants were instructed to identify what they considered to be irregularities in the images. For all sessions, images were displayed for 10 seconds before input from the user was accepted. For the gaze-directed groups, the SGD technique was applied during this time frame (but not during the subsequent marking stage). This helped to minimize the impact of viewing duration on the results and also allowed enough time to guide the viewers gaze about the images. Ten seconds was chosen based on the experts average viewing time before making the first selection (9.71 seconds).

The pairs of images for each patient were arranged so that the image of the left breast was placed on the right and the image of the right breast was placed on the left to mimic the preferred arrangement of the expert. The eyetracker used in this study is a SensoMotoric Instruments iView X Remote Eye Tracking Device operating at 250 Hz with gaze position accuracy < 0.5°. A chin rest was not used in this study.
3.4.1.2 Participant

20 novice participants (7 females, 13 males), between the ages of 18 and 30 volunteered to participate in this study. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. Participants were randomly assigned to one of four groups:

- **Static group:** 5 participants were presented with a randomized sequence of the 20 training images without the use of gaze manipulation. This group served as the control group for the experiment.

- **Gaze-directed group using expert scanpath:** 5 participants were presented with a randomized sequence of the 20 training images with gaze manipulation used to guide them to follow a similar scanpath (sequence of fixations) as the expert.

- **Gaze-directed group using expert selections:** 5 participants were presented with a randomized sequence of the 20 training images with gaze manipulation used to guide them only to the regions marked by the expert as irregularities. The overall scanpath of the expert was not used.

- **Gaze-directed group using adversarial scanpath:** 5 participants were presented with a randomized sequence of the 20 training images with gaze manipulation used to guide them along a scanpath that was chosen by the researchers to follow different directions and locations than that of the expert.

We hypothesized that using gaze manipulation to guide a novice along a similar scanpath as the expert or directly to the locations marked by the expert would improve the likelihood in correctly identifying irregular regions on the mammograms. Similarly, we
hypothesized that using gaze manipulation to guide the novices focus along an adversarial scanpath would reduce the likelihood that the novice correctly identifies the irregular regions on the mammograms.

3.4.1.3 Procedure

The expert radiologist hired to participate in this study helped to guide the user interface design of our training platform so that it closely mimics that of standard digital mammography systems. The expert also described external factors such as preferred lighting conditions for studying mammograms which we incorporated into the design of our experiment. The experiment was divided into four stages: a tutorial session, a training session, and two followup sessions to determine if the novice users had become sensitized to the training method used.

- Tutorial Session:

  All participants took part in the tutorial session. They were each presented with the same five images in a randomized order and asked to identify what they considered to be irregularities in the images. The images were displayed for 10 seconds before user input was accepted. A left mouse click-and-drag motion was used to draw a circular region that enclosed the abnormality. The Shift key was locked in software to ensure that the region drawn was always circular. Multiple regions could be selected in this manner. A double-click moved on to the next image. The sole purpose of the tutorial session was to get users familiar with the user interface and method of region selection to be used in the remaining stages of the experiment. No feedback about the accuracy of their selected regions was given.
Figure 3.17: Experts raw scanpath (a), simplified expert scanpath (b), adversarial scanpath (c), and comparison of simplified expert scanpath and adversarial scanpath (d) for one image from the dataset. The large purple circle shows the area marked by the expert as an irregularity. The numbers in the smaller circles of (a) indicate the order of fixations of the expert. The numbers in the smaller circles of (b) and (c) indicate the order of modulations used to guide participants in the corresponding groups.

- **Training Session:**

All participants took part in the training session. They were presented with the same twenty images in a randomized order and asked to identify what they considered to be irregularities in the images. For participants in the gaze-directed groups, the SGD technique was used to influence the order and location of their fixations during the first 10 seconds of viewing. The SGD approach proposed in section 3.1 was used to modulate the subtle visual cues.

To determine the location and order of the modulations for the expert-scanpath gaze-directed group, the experts scanpath was replayed in slow motion and the researchers manually selected what they considered to be the center of each cluster of fixations in sequence. This manual approach worked well since there were only 20 expert videos to consider. Of course, clustering algorithms [119] or connected
components analysis [120] could be used to automate this process. Using these simplified expert scanpaths as a guide, the researchers also manually selected the order and locations of the modulations for the adversarial scanpaths. These were chosen to follow different directions and locations than that of the expert. Figure 3.17 shows an example of the experts raw scanpath and the resulting simplified scanpath and adversarial scanpath. It also shows an overlay of both the simplified scanpath and the adversarial scanpath to illustrate how different they are. For the expert-selection gaze-directed group, the modulations corresponded to the centers of the regions marked by the expert as containing irregularities. The modulations were presented in the order they were marked.

- **Short-Term Follow-up Session:** Following the training session, the participants were given a five minute break. A follow-up session was run using twenty fresh images. No gaze manipulation was used in this session for any of the participants. The purpose of this follow-up session was to determine if the novice participants had become sensitized (short-term) to the strategies used during the training session.

- **Long-Term Follow-up Session:** The participants were asked to return one week later for a second follow-up study. Similar to the first follow-up session, this session was run using twenty fresh images and no gaze manipulation was used for any of the participants. The purpose of this follow-up session was to determine if the novice participants had become sensitized (long-term) to the strategies used during the training session.
3.4.2 Results and Discussion

To facilitate analysis of the novices’ performance, we define a weight-based accuracy measure as follows:

\[
\text{Accuracy} = \left( \frac{1 + (w_h \ast h + w_c \ast c + w_m \ast m)}{1 + h + c + m} \right) \ast 100
\] (3.8)

where \( h \) is the number of hits, \( c \) is the number of close matches, and \( m \) is the number of misses. \( w_h, w_c, \) and \( w_m \) are corresponding weights. Figure 3.18 illustrates the concepts of hits, close matches, and misses which are defined in terms of the circles drawn by the expert and the circles drawn by the novice. A hit occurs if the distance between the centers is less than either of the radii as shown in the top row of Figure 3.18. A close match occurs if the distance between the centers of the circles is less than the sum of the radii as shown in the bottom left of Figure 3.18. Finally, a miss occurs if the circles do not overlap as shown in the bottom right of Figure 3.18. If multiple novice selections result in hits to a single region selected by the expert, only one is considered to be a hit and the others are ignored. Likewise if multiple novice selections result in close matches to a region selected by the expert, only one is considered to be a close match and the others are ignored. In the case where there is a mix of hits and close matches to a single region selected by the expert, only one hit is considered and the others are ignored. We assign the following weights: \( w_h = 1, \ w_c = 0.75, \) and \( w_m = 0. \) Note that the weight for a close match is biased more towards a hit than a miss. The reason for this is that the selection is sufficiently close to an abnormality to warrant further investigation instead of labeling it simply as a random selection.
Figure 3.18: Illustration of hits, close matches, and misses. The circles represent selections by the expert and novice. The center of the circles are represented by the crosses.

<table>
<thead>
<tr>
<th>Classification</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive (TP)</td>
<td>Abnormal regions exist, and are correctly identified by the subject</td>
</tr>
<tr>
<td>False positive (FP)</td>
<td>No abnormal regions exist, yet subject marked regions as being abnormal</td>
</tr>
<tr>
<td>True negative (TN)</td>
<td>No abnormal regions exist, and no regions were marked by subject as being abnormal</td>
</tr>
<tr>
<td>False negative (FN)</td>
<td>Abnormal regions exist, but were not marked by the subject</td>
</tr>
</tbody>
</table>

Table 3.2: Mapping of test outcomes to true positives, true negatives, false positives, and false negatives.

Alternatively, binary classification statistics [121] [122] can be used to establish measures of accuracy as well as sensitivity and specificity. To calculate these properties it is necessary to categorize the test outcomes as true positives, true negatives, false positives, and false negatives. Table 3.2 shows the mapping of test outcomes to these conditions.

For this study sensitivity refers to the ability of the subject to correctly identify existing abnormalities in the mammograms. Sensitivity is computed as follows:
Sensitivity = \frac{\text{# of TP}}{\text{# of TP} + \text{# of FN}} \times 100 \quad (3.9)

On the other hand, Specificity refers to the ability of the subject to correctly identify mammograms containing no abnormalities. Specificity is defined as follows:

Specificity = \frac{\text{# of TN}}{\text{# of TN} + \text{# of FP}} \times 100 \quad (3.10)

The sensitivity and specificity values can then be combined to produce a binary classification-based measure of accuracy as follows:

Accuracy = \frac{\text{# of TP} + \text{# of TN}}{\text{# of TP} + \text{# of TN} + \text{# of FN} + \text{# of FP}} \times 100 \quad (3.11)

3.4.2.1 Training Session Results

Figure 3.19 shows the average weight-based accuracy (equation 3.8) for the various groups of participants during the training session. The static group averaged 54.2\%, the group that was guided by the expert scanpath averaged 64.5\%, the group that was guided by the expert selections averaged 68.8\%, and the group that was guided by the adversarial scanpath averaged 55.5\%. These values were obtained by averaging the accuracy for all participants in a group over all images used in the training session.

The averages show that the participants from both expert guided groups were more accurate than the participants from the static group. Independent-samples t-tests reveal that this effect was significant in both cases and not due to chance:

\[ t(198) = 3.227; p < 0.05 \quad (\text{expert scanpath vs. static}) \]
Figure 3.19: Average weight-based accuracy (equation 3.8) for different groups during the training session. The error bars represent standard error.

Figure 3.20: Binary classification-based accuracy (equation 3.11) for different groups during the training session.

\[ t(198) = 4.530; p < 0.05 \ (\text{expert selection vs. static}) \]

Figure 3.19 shows the binary classification-based accuracy (equation 3.11) for the various groups of participants during the training session. The static group averaged 53.6%, the group that was guided by the expert scanpath averaged 68.3%, the group that was guided by the expert selections averaged 64.7%, and the group that was guided by the adversarial scanpath averaged 52.0%. When using this accuracy measure we again observe that both expert-guided groups perform better than the static group.

These observations are not surprising as previous studies have already established that guiding attention to the relevant regions of a scene facilitates task completion. With
SGD however, there is the added benefit that the cues used to attract the viewer’s attention have minimal impact on the viewing experience as they occur only in the viewer’s peripheral vision and do not permanently alter the overall appearance of the image being viewed.

3.4.3 Short-Term Follow-up Session Results

Figure 3.21 shows the average weight-based accuracy (equation 3.8) for the various groups of participants during the short-term follow-up session. The static group averaged 51.3%, the group that was guided by the expert scanpath during the training session averaged 59.5%, the group that was guided by the expert selections during the training session averaged 63.1%, and the group that was guided by the adversarial scanpath during the training session averaged 53.9%. These values were obtained by averaging the accuracy for all participants in a group over all images used in the short-term follow-up session.

The averages show that the participants who were trained using the expert-guided approaches performed better than participants who were not gaze directed during training.
This indicates that there are short-term lingering effects related to the use of gaze manipulation and that the participants are becoming sensitized to the method of training used. Independent-samples t-tests confirm that this effect was significant in both cases and not due to chance:

$$t(198) = 2.311; p < 0.05 \ (expert \ scanpath \ vs. \ static)$$

$$t(198) = 3.189; p < 0.05 \ (expert \ selection \ vs. \ static)$$

Figure 3.22 shows the binary classification based accuracy (equation 3.11) for the various groups of participants during the short-term follow-up session. The static group averaged 26.9%, the group that was guided by the expert scanpath averaged 45.8%, the group that was guided by the expert selections averaged 41.6%, and the group that was guided by the adversarial scanpath averaged 27.5%. When using this accuracy measure we again observe that both expert-guided groups continued to perform better than the static group in the short-term follow-up session.
3.4.3.1 Long-Term Follow-up Session Results

Figure 3.23 shows the average weight-based accuracy (equation 3.8) for the various groups of participants during the long-term follow-up session. The static group averaged 64.2%, the expert scanpath group averaged 64.2%, the expert selection group averaged 65.4%, and the adversarial scanpath group averaged 65.2%. These values were obtained by averaging the accuracy for all participants in a group over all images used in the long-term follow-up session. The averages show that there is little difference between the groups and independent-samples t-tests confirm that the differences are not significant. This suggests that there are no long-term lingering effects related to the use of gaze manipulation.

Figure 3.24 shows the binary classification-based accuracy (equation 3.11) for the various groups of participants during the long-term follow-up session. The static group averaged 51.6%, the group that was guided by the expert scanpath averaged 59.1%, the group that was guided by the expert selections averaged 53.0%, and the group that was guided by the adversarial scanpath averaged 51.5%. When using this accuracy measure we again observe that there is little variation between the groups with the exception of the expert scanpath group which appears to be higher. Further analysis using the receiver
operating characteristics (ROC) space (described below) shows however, that all of the groups are actually performing at the accuracy level equivalent to a random guess.

### 3.4.3.2 Receiver Operating Characteristic (ROC) Results

Figure 3.25 shows the sensitivity and specificity for each group of participants across all sessions. Using these values we can generate receiver operating characteristic (ROC) plots for each session as shown in Figure 3.26. ROC plots (i.e. sensitivity vs. 1-specificity) help us assess the performance of the training techniques used. The blue dotted line at 45° represents the line of random guess. Any data above this line indicates a better-than-random classification and below this line a worse-than-random classification.
Figure 3.26: ROC plots for each session of the experiment for all groups. From left to right: training session, short-term follow-up session, and long-term follow-up session.

The ROC plot for the training session shows that the group guided by expert selection had a higher specificity compared to the other groups. The higher specificity may be attributed to the fact that the subjects associated the presence of the subtle modulations with the existence of abnormalities. Consequently when no modulations were present they were less likely to mark the image as being irregular. The ROC plot for the training session also shows that both expert-guided groups were more sensitive than the static and adversarial groups. This supports the accuracy findings described above.

The ROC plot for the short-term follow-up session shows that the expert-guided groups perform better than the control group and the adversary group. Note however that both the sensitivity and specificity for all groups is lower than in the training session. This suggests that there was some common phenomenon affecting the subjects of all groups. We believe that this could be due to fatigue or boredom (recall that the participants had already viewed 25 images lasting a minimum of 10 seconds each - not including the time it took for them to respond to each image).

The ROC plot for the long-term follow-up session shows that the performance of all the groups lie on the random guess line indicating that there are no difference between the groups and that there are no long-term lingering effects associated with the training.
methods used. Notice also that the phenomenon that caused the lower sensitivity and specificity observed during the short-term follow-up session is no longer present after a week-long break.

3.4.4 Conclusion

In this section we explored whether subtle gaze manipulation is capable of improving the efficiency of mammography training. Using a weight-based accuracy measure we observed the following:

- For the training session, participants in both expert-guided groups performed significantly better than the participants in the static group.

- For the short-term follow-up session, participants in both expert-guided groups continued to perform significantly better than the participants in the static group.

- For the long-term follow-up session, no significant effects were observed.

Similar results were observed using the binary classification accuracy measure for all sessions. The general agreement between these two accuracy measures provides added assurance about the reliability of the results.

Based on the ROC plots we can conclude that actively guiding the subjects using the expert-based approaches results in better performance in terms of sensitivity and specificity. In the short-term follow-up session we noticed that both sensitivity and specificity of the group guided by the expert selections dropped dramatically compared to the group guided using the expert scanpath. For this reason, and for the fact that constantly modulating a single selection may become annoying, we conclude that it is better to guide the novices using the expert scanpath.
Overall, our findings indicate that using the SGD technique to actively guide novice participants along the scanpath of an expert significantly improves accuracy compared to static viewing. In addition to exhibiting great promise for digital mammography training as well as training on other medical image modalities such as X-rays, CT scans, PET scans, and MRIs, these observations have implications for a wide range of visual search and learning applications. To test the application in another real-world scenario we conducted an experiment to see the impact of gaze guidance (Overt and Subtle) for password recollection.

3.5 Gaze Guidance for Improved Password Recollection

Recalling passwords has become a routine activity of modern life. Security of user information is critical due to the rapid increase of user accessible computing devices and cloud-based application interfaces. Many systems mandate that users create complicated passwords which are required to be a certain length and must include a combination of alphanumeric and special characters. For additional security, certain systems also require periodic password changes and the new passwords cannot be similar to previously used passwords. Shorter numeric passwords are also used widely for ATMs, locker systems, and mobile devices. Remembering all of these passwords is a very challenging task. Furthermore, the procedures for recovering/resetting passwords are time consuming and often lead to user frustration. The use of password managing software can help address these challenges however such technologies are not yet widely adopted.

Researchers have proposed many theories related to pre-attentive visual processing of information from the environment. The scanpath theory [123] suggests that successive saccades and fixations made while viewing an image become part of the lasting memory
Figure 3.27: Experiment Design. Images in the top and bottom rows are sample screen shots from the memorization and recall stages respectively. The columns indicate the four conditions namely, No-Image, Image-Only, Overt-Guidance, and Subtle-Guidance. Each image in the middle row is a screen shot containing 18 random 7-digit numbers from the cognitive task stage. In the memorization stage, for conditions 3 and 4, participants associate each digit in the password to regions in the image using a mouse. The selected regions are indicated by blue circles. For condition 3, during the recall stage, the participants are overtly guided to the previously marked regions in sequence using blue circles. For condition 4, the participants are guided to the previously marked regions using subtle image space modulations presented only in the viewer’s peripheral field of view. Inset image A shows a sub-image extracted around the password associated region selected by the participant. Inset images B and C show the bright and dark extents of the luminance modulations respectively.

and if replayed can lead to better scene recollection. Guided search theory [124] models the goals of the viewer’s search behavior and attempts to differentiate valid information from distractors. Researchers have also studied how image features influence visual attention and enable task completion [3]. It has also been shown that visual short term memory is highly dependent on the regions in the scene that attract visual attention [86].

This section explores if spatial visual cues can improve password recollection. Specifically we examine if associating each character of a password with a spatial region of an image and then guiding the viewer to revisit these regions leads to better recollection. Also known as the method of loci, this idea of associating items that one would like to remember with auxiliary spatial information has been shown to enhance memory recall [125]. Variations of this technique are commonly used by memory competition participants.
3.5.1 Experiment Design

We tested subjects’ ability to recall three randomly-generated numeric passwords under each of the following conditions:

- **Condition 1: No-Image** - during a memorization stage, participants were provided with three randomly generated passwords against a black background. In a subsequent recall stage participants were asked to recollect the three passwords.

- **Condition 2: Image-Only** - during a memorization stage, participants were provided with three randomly generated passwords along with a unique image for each password. In a subsequent recall stage, participants were shown the images and asked to recollect the corresponding passwords.

- **Condition 3: Overt-Guidance** - during a memorization stage, participants were provided with three randomly generated passwords along with a unique image for each password. Participants were asked to associate specific regions of the image with each digit in the password. In a subsequent recall stage, participants were shown the images and overtly guided to the marked regions in sequence and asked to recollect the corresponding passwords.

- **Condition 4: Subtle-Guidance** - during a memorization stage, participants were provided with three randomly generated passwords along with a unique image for each password. Participants were asked to associate specific regions of the image with each digit in the password. In a subsequent recall stage, participants were shown the images and subtly guided to the marked regions in sequence and asked to recollect the corresponding passwords.
Between the memorization and recall stages, the participants were presented with the cognitive task of reading 18 randomly generated 7-digit numbers out loud. Figure 3.27 illustrates each condition.

3.5.1.1 Stimuli

The stimuli images for all conditions were presented on a 22 inch monitor, operating at 60Hz with a resolution of 1680 x 1050. The images were obtained from the MIT benchmark images dataset [126]. Nine images (3 for each of the image-based conditions) were chosen by the researchers. The images contained multiple objects and salient regions to facilitate password association. All images were padded with a gray background to fill the screen. A SensoMotoric Instruments iView X Remote Eye Tracking Device operated at 120Hz, with a manufacturer-reported gaze position accuracy < 0.5°, was used in this experiment. A chin-rest was not used in this experiment in order to encourage natural viewing. The participants were seated with a viewing distance of ~70 cm from the eye tracker.

3.5.1.2 Participants

17 participants (4 females, 13 males), between the ages of 19 and 30 (avg. 24) volunteered to participate in this study. All participants reported normal and corrected-to-normal vision with no color vision abnormalities. Two participants’ data were removed from analysis due to tracking loss.
3.5.1.3 Procedure

Each participant was asked to recall a total of 12 passwords, three for each condition. New passwords were generated for each subject. The order in which the conditions were presented to the subjects was randomized. There was a 2-minute break between each condition to allow the researchers time to modify the experiment parameters. The participants viewed a blank screen during this time.

3.5.1.4 Memorization Stage

During the memorization stage, three unique randomly generated 7-digit passwords were shown to the participant consecutively. The participants were given unlimited time to attempt to memorize the password and had to click on a GUI button to move to the next one. For the Overt-Guidance and Subtle-Guidance conditions, they were asked to first associate a region of the image with each of the 7 password digits. This was done by clicking on the 7 desired regions in order. The clicked regions were highlighted using a blue circle and remained highlighted during memorization. Participants were not eye-tracked during the memorization stage.

3.5.1.5 Cognitive Task

Following the memorization stage, 18 random seven digit numbers were displayed on screen. The participants were asked to read them aloud. This was done to increase the difficulty of the recollection task. After reading the numbers, the participant clicked a button to move to the next stage. Participants were not eye-tracked while reading.
3.5.1.6 Recall Stage

At the beginning of the recall stage, a 9-point eye-tracker calibration was performed, followed by a 4-point validation using manufacturer provided techniques. The validation step revealed calibration accuracy for all participants to be $< 1^\circ$. The participants were asked to recall the 3 passwords from the memorization stage by entering them on three consecutive screens. For each screen, the participants were required to wait 25 seconds before entering their response. An audible beep was played after 25 seconds, and the keyboard and mouse were activated.

The 25 second delay was necessary to allow for the active guidance of participants’ gaze in conditions 3 and 4. The delay was applied to all conditions for consistency. For the gaze guided conditions, the previously marked regions were highlighted in order and cycled through three times. For the Overt-Guidance condition, this involved displaying each of the marked regions using a blue circle for 1-second (total 21 seconds). For the Subtle-Guidance condition, the timing is non-deterministic since modulations are dependent on the subject’s gaze. Modulations are presented only in the peripheral regions of the field of view and are terminated once a saccade is detected towards the target. We rely on earlier results indicating that this termination criteria is met within 0.5 seconds for approximately 75 percent of target regions and within 1 second for approximately 90 percent of target regions [67]. So the expectation is that 3 cycles of the password modulations would be completed within 21 seconds. We added 4 seconds to further increase the likelihood of cycle completion.

For the Subtle-Guidance condition, modulations occurred at a frequency of 60Hz and the intensity of the modulated pixels varied along a sinusoidal curve between $\pm 10\%$ of the original pixel intensity value multiplied by the corresponding value from a predefined
Figure 3.28: Average number of digits correctly recalled for each condition. Error bar indicates standard error.

Gaussian distribution. In addition to gaze information, the participants’ input password and response time after the beep were recorded.

3.5.2 Results and Discussion

3.5.2.1 Accuracy

The task of memorizing three new 7-digit passwords and attempting to recall them after reading eighteen 7-digit numbers out loud is obviously a difficult one, and our results reflect this fact. Using a simple binary (right/wrong) metric, the Overt-Guidance condition posted the highest number of correctly recalled passwords with an average accuracy of 36%. The least accurate was the No-Image condition, with an average accuracy of only 10%. The average accuracy of the Image-Only and Subtle-Guidance conditions were both 18%.

To obtain a more fine-grained accuracy measure, we then consider the average number of password digits recalled correctly and the average number of correct but misplaced digits.
Figure 3.28 shows the average number of password digits recalled correctly. As expected, this accuracy measure during the Overt-Guidance condition is significantly higher than the others. While not significant, the Subtle-Guidance condition also resulted in a higher average number of correct digits per password than the Image-Only and No-Image conditions. Figure 3.29 shows the average number of correct but misplaced digits. To prevent data from being skewed, we only count a misplaced digit once. This helps to minimize the effect of random guesses leading to a more accurate result. The Overt-Guidance condition posted significantly lower misplaced digits compared to the other conditions. The No-Image condition posted a significantly higher number of misplaced digits. A single-factor ANOVA was conducted between subjects to compare the four conditions. The results from the ANOVA indicated that these conditions were significantly different with p values < 0.05. These observations were further analyzed using independent-sample t-tests which are summarized in Table 3.3.
Table 3.3: Significant t-test Results

<table>
<thead>
<tr>
<th>Test Performed</th>
<th>t value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct Digits; No-Image vs. Overt</td>
<td>3.33</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Correct Digits; Image-Only vs. Overt</td>
<td>2.78</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Correct Digits; Subtle vs. Overt</td>
<td>2.52</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Misplaced; No-Image vs. Image-Only</td>
<td>1.61</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Misplaced; No-Image vs. Subtle</td>
<td>1.61</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Response Time; No-Image vs. Overt</td>
<td>2.21</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Response Time; No-Image vs. Subtle</td>
<td>2.43</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Response Time; Image-Only vs. Overt</td>
<td>1.98</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Response Time; Image-Only vs. Subtle</td>
<td>1.98</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Dwell-Time; Subtle vs. Overt</td>
<td>1.19</td>
<td>&lt; 0.05</td>
</tr>
</tbody>
</table>

Figure 3.30: Average time used to recall password for each condition. Error bar indicates standard error.

3.5.2.2 Response Time

Response time was measured from the onset of the system beep until the password entry was completed. The average response time across all conditions is shown in Figure 3.30. The average response time for the gaze guided conditions were significantly slower than that of the unguided conditions (see Table 3.3). One possible explanation for this could be the additional cognitive load of having to follow the visual cues on screen.
3.5.3 Dwell-time on Target Regions

Dwell-time is the duration for which a subject fixates on an area of interest (AOI). For the gaze guided conditions, we define AOIs to be within 3° of visual angle of the regions marked by the subject. We observed no significant difference in average dwell-time between the gaze guided conditions. Figure 3.31 shows example scanpaths and marked regions from one subject for the gaze guided conditions.

3.5.4 Conclusion

We analyzed the effect of gaze guidance on password recollection. Results indicate that guiding attention aids in password recollection, with Overt-Guidance having significantly higher accuracy compared to the other conditions. Response time indicates that participants took more time to recall passwords when there was gaze guidance. Dwell-time in target regions for the gaze guided conditions were similar, indicating there was no significant difference in the average fixation duration for subjects between the gaze guided groups. All these experiments indicate that subtle gaze guidance is indeed effective in aiding short-term memory, spatial information recall, sequencing tasks, learning
Figure 3.32: Annotated photograph of the real-world gaze manipulation setup. The current fixation is determined using a wearable eye-tracker. A projector is used to create a visual cue on another part of the scene in order to guide the viewer’s gaze to that location. This process can be repeated to guide the viewer over time.

and training, and password recollection. Most of these experiments are confined to digital imagery in a laboratory setup. The idea of guiding attention in real-world environments is compelling and equally challenging. In the next section an experiment is conducted to perform gaze guidance in a controlled real-world environment.

3.6 Guiding Attention in Controlled Real-World Environments

Gaze manipulation has high applicability in instructional environments. Numerous studies have been conducted to understand experts’ eye movements for specific tasks and to use their fixation sequence to guide novices during training. This has been done in tasks such as aircraft inspection [76] and optic disc examinations [77]. Gaze manipulation approaches have shown to be effective for computer-based spatial learning [71, 72], search
task completion [74], and medical training applications [80, 82].

The concept of guiding attention (pointing) in physical spaces is well established in art, photography theatre and cinema. In most cases this relies on clever use of contrast and composition, direct manipulation of lights, or manual post-processing to highlight regions or characters of interest. Several augmented reality systems have also been developed to directly highlight real-world objects [127–129] or to highlight objects on a video feed of the real world [130, 131]. However, none of these systems take the viewer’s gaze into account or attempt to guide attention without using overt cues. One notable exception is the work by Veas et al. [132] where they pre-process video feeds to increase the saliency of target regions while reducing the saliency of surrounding regions. They do this by adjusting the contrast of the various channels in CIELAB space to create a natural looking change to the image which increases the likelihood of the target regions being attended to. In their work, eye tracking was only used to test if the subjects actually paid attention to target regions. Furthermore, their technique does not actively try to shift attention between multiple targets.

The goal of this work is to develop and test various approaches to extend the concept of gaze manipulation beyond digital imagery to include controlled, real-world environments. In this system (see Figure 3.32), the user wears a pair of eye tracking glasses that allow for unrestricted head movement. The glasses are equipped with a front facing camera that captures the scene that the viewer is looking at and determines the viewer’s gaze position within the scene.

We extract the image features (SIFT) from this camera’s video feed near to the viewer’s current fixation and search for matching features in a corresponding high resolution source image. This allows us to infer the location of the viewer’s current fixation on the
high resolution image. Once we have established where the viewer is looking, a visual
cue (subtle luminance change or a more overt flash) can then be projected onto another
part of the scene to attract the viewer’s attention. This can be repeated as necessary
to guide the viewer’s gaze over time. We evaluated the effectiveness of our system by
conducting an experiment in which we guided viewers to look at sequences of objects in
a scene.

3.6.1 Experiment Design

The two main challenges to guiding attention in real world environments are (1) deter-
mining what the viewer is currently paying attention to, and (2) projecting visual cues
on other parts of the scene to draw the viewer’s attention. Our system combines the
use of eye tracking glasses, a standard projector, and a high-resolution source image
of the scene being viewed (taken approximately from the projector’s perspective). The
eye tracking system computes the fixation position within the scene that the viewer is
attending to. Scale Invariant Feature Transform (SIFT) descriptors are then extracted
in the regions surrounding the fixation point in the scene image. These are compared
against pre-computed SIFT descriptors from the source image. The output from this
matching algorithm allows us to infer the location of the viewer’s fixation on the source
image. This is an important step in our process as it contributes to the subtlety of
technique by ensuring that we do not attempt to guide the viewer to objects they are
already attending to. Once the current fixation is determined, viewer attention can then
be guided by projecting a visual cue onto another part of the scene.
The following design goals were established:

- The system should perform in real-time.

- The system should allow for unrestricted movement of the viewer as long as their fixations remain within the region covered by the source image.

- The system must be robust even under ill-conditioned and ill-posed situations such as track-loss or when the viewer looks away from the scene.

The following assumptions were made:

- The system will be used in a visually stable environment (i.e. unchanging external lighting and no moving objects).

- The objects in the scene are non-transparent and well defined in terms of color, contrast, size, shape and texture.

These assumptions are necessary to ensure a high correlation between the SIFT descriptors from the eye tracker scene camera and the SIFT descriptors from the source image. To accomplish all of this in real-time, our system performs image matching on a frame-by-frame basis using multiple GPUs.

3.6.1.1 System Overview

Figure 3.33 shows the architecture for our real-world gaze manipulation framework. The eye tracker used in our system is a SensoMotoric Instruments (SMI) Eye Tracking Glasses. The glasses are equipped with a front facing camera (scene camera) that captures the scene that the viewer is looking at as well as two rear facing cameras that
capture binocular eye movements (eye cameras). The resolution of the scene camera is 1280 x 960 @ 24p and the field of view is 60° horizontal and 46° vertical. Gaze position within the scene is computed using SMI’s proprietary eye tracking software at a rate of 30 Hz. The eye tracker is connected to a dedicated laptop which computes the viewer’s fixations on the scene camera frames in real time. The laptop is powered by a dual core processor with 4 GB RAM. The compressed video and fixation data are streamed over the network to a desktop for further processing. The desktop is equipped with 24 GB RAM, dual hexa-core processors and dual GTX 590 graphics cards. A standard RGB projector completes the hardware for our system.
SIFT descriptors of the source image are computed ahead of time using one of the GPU cores. The eye tracker system streams gaze positions and scene camera frames in real time over the network through a UDP socket. These gaze points and the compressed scene frames are added to a FIFO queue. Three GPUs are used to process this information to compute the SIFT descriptors. A dedicated GPU core performs SIFT descriptor matching between the scene and the source image. To reduce computation time and improve the likelihood of solution convergence, only a subset of the SIFT matches are extracted around the current fixation point. The resulting matching points are used by a multi-threaded CPU program to infer where the viewer is attending to in the source image. This distribution of work between the GPUs and CPU allows our system to run in real-time.

Our system was developed in C++ and utilizes a variety of libraries for specific tasks including SMI’s proprietary API for interfacing with the eye tracking glasses, Sift-GPU [133], the Eigen library [134] for linear algebra and related algorithms, and finally SFML for data visualization. Underlying technologies include CUDA and OpenGL for SiftGPU and OpenMP for the Eigen library.

### 3.6.1.2 SIFT descriptor extraction and matching

Sparse SIFT descriptors are computed for both the scene and source images using Sift-GPU. The thresholding values are set manually to obtain optimum results for the given scene. This only has to be done once since we are dealing with controlled environments where the thresholds will not change from frame-to-frame.

Once descriptors for the current frame have been identified, they are matched with the precomputed descriptors from the source image as shown in Figure 3.34. To filter
Figure 3.34: SIFT descriptor extraction and matching and fixation inference. Source image (top left), scene image (top right), and 5% of the resulting descriptor matches (bottom). Correct matches are shown in green and an incorrect match (outlier) is highlighted in magenta; current fixation is shown as a red cross; inferred fixation is shown as a green cross; the red circle shows the area where SIFT descriptors are matched.

the matching results, we perform a radial search about the fixation point until 5% of the total number of descriptors in the source image have been found. We limit this as it provides sufficient information to infer the fixation position while not incurring any additional computational costs. To eliminate incorrectly matched points from this subset of descriptors we use Cook’s Distance measure [135] which performs dimensionless weighting of all values in a set, with respect to the mean of that set. In our case, we expect that the positional average of the subset of descriptors obtained in the scene image will be close to the fixation point (in the Cartesian sense). This means that the matched points with higher Cook’s Distance values are more likely to be outliers. Once the outliers have been eliminated, a least squares solver is used to infer the fixation on the source image.
3.6.1.3 Stability Considerations

Occasionally, situations may arise that lead to inaccurate results or an inability to compute the fixation position on the source image. The most common is a track-loss (i.e. the eye-tracker is not able to detect the pupil due to blinks or extreme pupil position). In this case the scene video frames will have no accompanying gaze information making fixation inference impossible. When this occurs we simply discard the frames and resume once gaze positions are again available. Also, if the viewer happens to look away from the scene, then the inferred gaze position will naturally be incorrect. We detect when this is happening by keeping track of the number of matched descriptors from frame-to-frame. If the number changes drastically (indicating dissimilar scenes) then we ignore the results from the SIFT descriptor matching algorithm. Finally, it is possible (though highly improbable) for the least squares solver not to converge. This can only happen in cases where there are fewer than two descriptors present (i.e. fixation occurs in a large uniform region of the scene) or if all descriptors are co-linear. In these cases we simply reuse the previous frame’s result. By addressing all of these ill-posed and ill-conditioned situations, we ensure that the system is both accurate and stable.

3.6.1.4 User Study

The goal of the user study was to test the effectiveness of our real-world gaze manipulation technique. Participants viewed a simple scene consisting of eight objects. The intended viewing order was not prescribed. After viewing the scene for a short period of time, we then attempted to guide their attention through six sequences of objects in the scene. The relevant objects were highlighted by projecting a brief luminance modulation using the projector.
The modulations were constructed by alternately blending some amount of black, then some amount of white. The rate at which the blend is modulated is 10Hz. A Gaussian falloff is used to soften the edges of the modulated regions. In our viewing configuration, the modulation diameter on the physical objects ranged from approximately 2 centimeters to 4 centimeters (depending on distance from the projector).

Figure 3.34 (top left) shows the source image of the scene used in our user study. The objects in the scene were non-transparent and well defined in terms of color, contrast, size, shape and texture. The objects were highlighted in a randomized order to minimize the introduction of learning effects.

Twenty participants (16 males, 4 females) between the ages of 18 and 29 volunteered for the user study. They all had normal, or corrected-to-normal, vision and were naive to the purpose of the experiment. They were simply instructed to look at the scene. Each participant underwent a brief calibration procedure, away from the scene, to ensure proper eye tracking. Six randomly generated viewing sequences were presented to each participant with a 10 second gap between sequences. Each sequence consisted of 8 objects. Between sequences the participants were told to reposition themselves however they saw fit but to keep the scene visible. This ensured that the data collected across all subjects covered a wide range of vantage points. The entire experiment (including calibration) for each participant was less than 10 minutes. We recorded the viewer’s gaze positions within the scene as well as the individual scene camera frames during the experiment. Data from one participant was excluded due to an extended period of track loss.
3.6.2 Results and Discussion

We needed a robust mechanism to compare the intended viewing sequence with the actual viewing sequence of each participant. The actual viewing sequence is extracted from the eye tracking data by identifying the first fixation that occurs after the onset of the visual cue. Levenshtein distance [101] provides an appropriate measure to compare distances between ordered sequences, such as those recorded during our experiment. We assign labels A through H to the eight objects in the scene as shown in Figure 3.37 (inset) and compare the intended viewing sequence with the actual viewing sequence. Suppose for the eight objects in the scene that the correct viewing order is ABCDEFGH. A Levenshtein distance of 0 would indicate no difference, whereas a distance of 8 would indicate the maximum difference.

Figure 3.35 shows the average and standard deviation of the Levenshtein distance for the six sequences for all participants. The similar averages and standard deviation values indicate that participant performance remains consistent over time. The average of the Levenshtein distance across all sequences for all participants is 0.85 (recall that the Levenshtein distance measure for this study ranges from 0 to 8). This means that we can typically expect less than one error per sequence. The histogram in Figure 3.36 shows the distribution of Levenshtein distance for all sequences. Notice that it is skewed to the left indicating that a large number distance measures are close to zero. Only 18% of the trials had more than one sequence error and the error rate falls off rapidly.

We also observed that the response time between the onset of the modulation and the subsequent fixation on the object was approximately 0.5 seconds. This is consistent with what was observed in the SGD technique for digital images.
Overall these results indicate that our technique is indeed effective at guiding attention in simple controlled real-world environments. Figure 3.37 shows a representative frame from the scene camera video that was captured from one subject for one of the sequences. It is overlaid with the subject’s scanpath. The numbered red circles indicate the order of the fixations that occurred during the presentation of the sequence and the size of the circle indicates the duration of the fixation. In this particular example there was a perfect match between the target sequence and the actual sequence.

To demonstrate the usefulness of real-world gaze manipulation, we used our framework in a parts retrieval experiment. Parts from a toy building set were arranged on a table.
in piles (see Figure 3.38) and we asked six subjects to retrieve 10 parts. Three subjects were given a sheet of paper with photos of the 10 parts (control) and three subjects were told that we were going to attempt to guide them to the correct parts. The results of this experiment are shown in Figure 3.38. As expected, subjects in the control group had to shift their attention back and forth between the printed sheet and the parts on the table. They also had to develop a strategy for keeping track of which parts were already found. This resulted in longer completion times and greater likelihood of error. On the other hand, the gaze guided group performed much better in terms of completion time and error rate due to reduced cognitive load.
3.6.3 Conclusion

The work presented in this section extends the SGD technique to guide viewer attention in controlled real-world environments. By projecting subtle luminance modulations into the physical world, our system is capable of drawing attention to a target region very quickly (response time $\approx 0.5$ seconds) and the process can be repeated on other parts of the scene to guide the viewer attention over time. Results of a user study reveal that our approach effectively guides viewers through sequences of objects with less than one error per sequence consisting of eight objects. Furthermore, the likelihood of more than one error is only 18%.

This technique has high applicability in instructional environments such as aircraft cockpits. For example, it can be used to direct novice pilots to pay attention to situation-relevant instrumentation in response to an event such as equipment failure or adverse weather conditions or to help them learn the order of operational protocols such as take-off and landing procedures. With our system, it is also possible to capture the fixation sequence of veteran pilots and use it to guide novices. Similar training can be done in driving simulators and systems for navigation aid. An obvious next-step for our work is to experiment with more complex and dynamic environments. We will be able to leverage recently published work from the eye tracking community which documents
‘best practices’ for collecting eye tracking data in outdoor environments [136, 137] as well as work from computer vision and augmented-reality on object detection and tracking [138, 139]. In summary, we explored extending the SGD technique in guiding a viewers’ attention to physical objects in a controlled real world setting.

The main limitation of our setup is restricted mobility due to the dependence on a fixed projector. It would be better to present the visual cues on a head-mounted display, however, commercial eye-trackers with head-mounted displays for augmented reality applications are not yet commonplace. Note that other researchers have attempted to track the eyes from below/above the frame of head-mounted displays with limited success. Transitioning to an integrated system would be simple and will greatly improve mobility and also be less distracting to bystanders. An other important limitation of the system is that, the modulation regions where the subject’s gaze is to be guided was pre-computed manually and the sequence of regions are fixed ahead of time. This approach is both time consuming and cumbersome. This bottle-neck can be solved if a model can be developed that can automatically predict target regions for gaze guidance. Chapter 4 presents a model of human visual attention prediction that uses saliency maps, scene feature maps and task-based eye movements to predict regions of interest.
Chapter 4

Task-Based Eye Movement Prediction

Predicting gaze behavior of a human in a given scene is a very challenging task. There are multiple factors that influence human gaze behavior. The salient features in the scene, the task at hand and prior knowledge of the scene are some of the factors that highly influence gaze behavior. Visual saliency based models predict regions of interest that attract the gaze of a subject based on image features such as contrast, color, and orientation [3, 25, 39, 46]. There are other top-down computational models that combine saliency maps and scene context. Some top down models use face detection, object detection, and image blobs with visual saliency to gather visual attention details in the scene [28, 35, 140]. The task being undertaken has a very strong influence on deployment of attention [5]. We look for things that are relevant for the current task and pay less attention to irrelevant objects in the scene.

Researchers have shown that there is a high correlation between visual cognition and eye movements when dealing with complex tasks [141]. When subjects are asked to perform
a visually guided task their fixations were found to be on task-relevant locations. This finding was established using the “block-copying” task where the subjects were asked to assemble building blocks, and it was shown that subjects’ eye movements reveal the algorithm used for completing the task [142]. Others have studied gaze behavior while performing tasks in natural environments such as driving, sports, walking, etc [6, 56, 141, 143]. There have been many computational models using Bayesian approaches to integrate top-down and bottom-up salient cues [144]. Many gaze prediction algorithms have been proposed based on image scene features and visual saliency maps [145, 146].

Section 4.1 presents a model of human visual attention prediction that uses saliency maps, scene feature maps and task-based eye movements to predict regions of interest. The model is evaluated by studying the impact of subtle gaze guidance on spatial information recall on the predicted target regions. The predicted regions are compared with ground truth data determined by researchers to establish the correctness and accuracy of the automatically generated targets. Such a system is highly scalable, and can be used for different task-based gaze guidance applications.

4.1 Automatic Target Prediction and Subtle Gaze Guidance for Improved Spatial Information Recall

To gain a better understanding of the visual processing involved at the target regions, a study was previously conducted to determine the impact of SGD on short-term spatial information recall (section 3.2). Participants viewed a randomized sequence of images. Following each image, they were presented with a blank screen and asked to recall the location of specific objects. They were instructed to use the mouse to draw the smallest rectangles that bounded each target region. Their input was analyzed to determine
how accurate their short-term spatial recollection was in terms of number of targets, location, and shape. Significant performance improvements were observed in target count and location recall compared to a control group who viewed the images without guidance. No effect was observed on the recollection of target shape (measured in terms of the aspect ratios of the bounding boxes).

We hypothesize that no effect on shape recollection performance was observed because the investigators applied the modulations only to the center of the rectangle bounding the target region. To test this idea, we repeat the experiment using the same images and targets as mentioned in the section 3.2. In addition to the center of the bounding rectangle, we modulate a rough outline of the target object obtained using an edge map composed from a dyadic pyramid of low spatial frequency maps of the original image. With this approach, we observed a significant improvement in shape recollection as well as target count and location recollection.

In this section, we also address another limitation that was common among many of the gaze manipulation techniques we reviewed. In most cases, the target regions of the image were manually pre-selected by the researchers. This process has to be repeated for each stimulus image and quickly becomes tedious. To fully exploit the benefits of gaze manipulation strategies, automated target selection techniques are needed. One common approach is to use annotated image databases such as labelme [147] which have been previously labeled by others, however significant manual effort is still necessary to build these databases. Other approaches rely on image segmentation [148, 149] or object detection [150, 151] to identify potential targets in the image, however with these approaches, the resulting targets are not necessarily the most relevant or prominent ones.
The saliency of particular regions or objects in an image plays a vital role in guiding visual attention. Many visual saliency models have been proposed to predict where humans focus their attention in an image [3, 38, 48, 152]. These models can be used to provide a prioritized list of targets for guiding visual attention. These models are particularly useful in cluttered environments where many salient regions compete for the viewer’s attention.

In this section, we combine a traditional saliency map with an edge map, and image feature maps in order to facilitate both target prediction and better shape recollection. Using this approach we were able to automatically identify 79% of the target regions that were previously manually selected in the study mentioned in section 3.2. In a new study with a different set of images and recollection tasks, our approach correctly predicted 85% of the relevant regions (an average of 81% across both studies). These results are remarkable as the algorithm has no prior information of the task being assigned to the viewer.

4.1.1 Automatic Target Prediction Framework

Our target prediction framework is illustrated in Figure 4.1. The output of this framework is a modulation map which not only identifies the centers of the target regions but also includes an edge map which captures a rough representation of the shape of the target regions. This information is ultimately used to guide the viewer attention and facilitate better shape recollection.

For each stimulus image (A), an edge map, saliency map, and feature maps are first computed as follows:
Figure 4.1: Framework for automatically identifying regions of interest and shape boundaries for gaze guidance. An edge map, saliency map, and feature maps of the original image are combined to obtain the final modulation map. Source image courtesy of [153].

- **Edge Map**: The edge map (E) is computed by thresholding the low spatial frequency dyadic pyramid (B) of the original image. Each image in the pyramid is computed at 4 cycles/image down-sampled by a factor of 2 from the original image. This ensures that only strong edges are captured in the resulting edge map. We use the canny edge detection algorithm [154] to compute the edge for each level in the pyramid.

- **Saliency Map**: The saliency map (C) is computed using the algorithm proposed by Itti and Koch [34] using normalized center surround conspicuity maps obtained from image intensity, color (RGB) and Gabor orientated filters [155] (0°, 45°, 90° and 135°).
- **Feature Maps**: The feature maps (F1, F2, and F3) are generated by computing n-key features from the original image and three corresponding high spatial frequency images (D). The high spatial frequency images are obtained by applying high-pass filters to the original image (A). We use SIFT [156], SURF [157] and MSER [158] to compute the feature maps. For each feature detection algorithm the strongest 50 features were chosen. SIFT and SURF feature detection algorithms are widely used for object recognition and tracking. We use both to leverage the strengths of each. The MSER algorithm on the other hand is widely used as a method for blob detection and provides shape descriptors for objects in images. A linear combination of MSER with SIFT and SURF provides a robust composite feature map (F).

A weighted combination of the composite feature map (F) and the saliency map (C) results in a saliency + feature map (G). The centers of each region in the saliency + feature is used as a target location for guiding viewer attention. This is combined with the edge map (E) to generate the final modulation map (H). The connected edges surrounding the target center provide a rough estimate of the shape of the target object.

For gaze guidance, the edges and the center of each region are modulated using the subtle gaze direction approach. The center of the object is modulated using a faint circular luminance pulse with maximum intensity in the middle and fading gradually as the radius increases (2D Gaussian distribution). On the other hand, the connected edges are modulated using intensities computed from an enclosing circle (also with a 2D Gaussian distribution) with maximum intensity at the circumference and fading gradually towards the center as shown in Figure 4.2. Note that only the pixels on the edge are modulated, not the entire circle. As with the SGD technique, these modulations
only occur on regions in the peripheral vision as determined by real-time eye tracking. The modulations are terminated before they can be scrutinized by the viewer’s high acuity foveal vision.

4.1.2 Experiment Design

We conducted two independent experiments using our edge-modulated SGD framework:

- **Experiment 1:** In experiment 1, we repeat the study conducted in section 3.2 using the same images and targets selected by the researchers. The goal of this study was to determine if our approach had any effect on the recollection of target shape. In addition to modulating the centers of the pre-selected targets, we also modulate the rough outline of the target shape computed using our framework. We refer to this as “SGD w/shape modulation” in the rest of this section.

- **Experiment 2:** In the second experiment, using a different image dataset, we eliminate the manual selection of target regions by the researchers altogether and instead utilize our framework to select the target regions. Naturally, under these conditions, it is possible that our technique may select some target regions that are
unrelated to the recollection task. The goal of this experiment is to determine if these false positives adversely impact the viewer’s performance on the recollection task.

4.1.2.1 Stimuli

The stimulus images for both experiments were presented on a 22 inch wide screen monitor, operating at 60 Hz with a resolution of 1680 x 1050.

- **Experiment 1:** The images for experiment 1 were obtained from the study mentioned in section 3.2. They consisted of 28 images (3 training images and 25 test images) compiled from various sources. The images ranged from simple scenes with a few objects to complex scenes with many objects. The number of objects or regions that the participants were asked to recall for each image ranged from 1 to 9. The researchers used Miller’s observation [87] that the average human can only hold $7 \pm 2$ items in working memory to establish the upper limit of 9 for the experiment.

- **Experiment 2:** The stimuli for experiment 2 consisted of 25 images compiled from the MIT benchmark images dataset [126]. The number of objects or regions that the participants were asked to recall for each image ranged from 1 to 7.

4.1.2.2 Participants

20 participants (5 females, 15 males), between the ages of 19 and 28 (avg. 23) volunteered to participate in this study. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. Participants were randomly assigned to one of two groups:
• **Static group:** 10 participants were presented with a randomized sequence of the 50 test images (experiment 1 and 2) without the use of gaze manipulation. This group served as the control group for both experiments.

• **Gaze-directed group:** 10 participants were presented with a randomized sequence of 50 test images (experiment 1 and 2) with gaze manipulation turned on.

For experiment 1 we hypothesized that using SGD w/shape modulations to guide the viewer’s focus to the correct target regions would lead to improved count, location and shape recollection.

For experiment 2 we hypothesized that automatically generating targets for guidance would not adversely impact performance on the spatial recollection task in spite of the false positive target regions that might be generated. This hypothesis is formulated based on observations in other studies involving SGD where the presence of “distractors” (i.e. modulations at incorrect regions) helped to spread the viewer’s gaze across images and still led to improved performance on a search task compared to static viewing [74].

### 4.1.2.3 Procedure

From the subject’s perspective, the procedure for both experiments is identical. This allowed us to capture their responses for both image sets in one session.

Following the approach mentioned in section 3.2, each image was shown for 10 seconds and was then replaced with a blank screen. While the blank screen was being displayed, an audio question about the spatial content of the image was played. The questions were recorded in a normal voice by a male native English speaker. Sampling rate for
the audio questions was 44.1 kHz. This approach is preferred to displaying the question as text on-screen as this may disrupt the participant’s short-term memory of the image [88].

The participants responded to the questions by using the mouse to draw rectangular regions on the blank screen. Three images from the complete set were used in a brief training session to ensure that the participants understood the procedure for completing the experiment. During the training session, participants were encouraged to ask questions and were able to view their solution and the correct solution after each image (see Figure 4.3).

The gaze of subjects in the gaze-directed group was monitored in real time using a remote eye tracking device. This was done to ensure that modulations were only presented to the peripheral regions of the field of view and were immediately terminated as the viewer’s focus approached the modulated regions. The eye-tracker used in this study
is a SensoMotoric Instruments iView X Remote Eye Tracking Device operating at 120 Hz with gaze position accuracy < 0.5°. Although not required for this particular eye tracking system, a chin rest was used to ensure minimal head movement and the viewing distance was fixed at ≈ 65cm from the display. Modulations occurred at a rate of 60Hz. During modulation, the intensity of the modulated pixels varies along a sinusoidal curve between ±10% of the original pixel’s intensity value multiplied by the corresponding value from the relevant Gaussian distribution.

Section 3.2, defines three measures for spatial recall accuracy (counting error, location error, and shape error) that we also utilize in this section. For the sake of brevity the accuracy measures are not repeated in this section.

4.1.3 Results and Discussion

For both experiments we measured the impact of gaze manipulation (modulating both target centers and rough edge outline) on short-term spatial information recall by computing the participants counting error, location error, and shape error. In summary, We observed the following effects:

- SGD w/shape modulation using pre-selected targets, that are relevant to the recollection task, results in a significantly lower counting error, location error, and shape error compared to static viewing.

- Our framework for automatically selecting target regions correctly predicted 79% of the task relevant regions for the images in experiment 1 and 85% of the task relevant regions for the images in experiment 2 (an average of 81% across both experiments). This means that approximately 20% of the predicted regions are false-positives that are not related to the recollection task at hand. Even in the
Figure 4.4: Average counting error for participants from the static and gaze directed groups for experiment 1 (top) and experiment 2 (bottom). Error bars represent standard error.

presence of these false positives, significantly lower counting error, location error, and shape error were observed compared to static viewing.

4.1.3.1 Counting Error

Figure 4.4 shows the average counting error of the participants for both experiments. These values were obtained by averaging the counting error for all participants in a
group over all the images in each experiment.

For experiment 1, counting error for the static viewing group averaged 1.288 targets while the counting error for the gaze-directed group averaged 0.86 targets (for comparison, the averages from section 3.2 were 1.488 targets and 0.76 targets respectively). The differences in the averages show that SGD w/shape modulation at pre-selected relevant regions results in a lower counting error compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

$$t(476) = 2.6659; p < 0.05$$

For experiment 2, counting error for the static viewing group averaged 0.4625 targets while the counting error for the gaze-directed group averaged 0.268 targets. Note that the average counting error is lower in experiment 2 because the average number of targets per image is lower than that of experiment 1. The differences in the averages show that SGD w/shape modulation at the automatically generated targets results in a lower counting error compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

$$t(491) = 1.9354; p < 0.05$$

Figure 4.5 shows how the counting error varies as the number of regions the participants were asked to recall increases for experiment 1. As expected, the counting error increases as the recall task becomes more difficult. Note, however, that the counting error for SGD w/shape modulation is consistently lower than that of static viewing. These results are consistent with the results from section 3.2. A similar effect was also observed in
Figure 4.5: Average counting error for participants from the static and gaze directed groups as number of target regions increase for experiment 1.

4.1.3.2 Location Error

Figure 4.6 shows the average location error of the participants for both experiments. These values were obtained by averaging the location error for all participants in a group over all the images in each experiment.

For experiment 1, location error for the static viewing group averaged 128 pixels while the location error for the gaze-directed group averaged 88 pixels (for comparison, the averages from section 3.2 were 134 pixels and 99 pixels respectively). The differences in the averages show that SGD w/shape modulation at pre-selected relevant regions results in a lower location error compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

\[ t(498) = 5.5145; p < 0.05 \]
Figure 4.6: Average location error for participants from the static and gaze directed groups for experiment 1 (top) and experiment 2 (bottom). Error bars represent standard error.

For experiment 2, location error for the static viewing group averaged 121 pixels while the location error for the gaze-directed group averaged 78 pixels. The differences in the averages show that SGD w/shape modulation at the automatically generated targets results in a lower location error compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

\[ t(377) = 4.4954; p < 0.05 \]
Figure 4.7 shows how the location error varies as the number of regions the participants were asked to recall increases for experiment 1. The location error increases as the recall task becomes more difficult. Note, however, that the location error for SGD w/shape modulation is consistently lower than static viewing. A similar effect was also observed in experiment 2.

4.1.3.3 Shape Error

Figure 4.8 shows the average shape error of the participants for both experiments. These values were obtained by averaging the shape error for all participants in a group over all the images in each experiment.

For experiment 1, shape error for the static viewing group averaged 125 pixels while the shape error for the gaze-directed group averaged 89 pixels (for comparison the averages from section 3.2 were 132 pixels and 120 pixels respectively). The differences in the averages show that SGD w/shape modulation at pre-selected relevant regions results in a lower shape error compared to static viewing. An independent-samples t-test revealed

![Average location error as number of target increases (Experiment 1)](image)

**Figure 4.7:** Average location error for participants from the static and gaze directed groups as number of target regions increase for experiment 1.
Figure 4.8: Average shape error for participants from the static and gaze directed groups for experiment 1 (top) and experiment 2 (bottom). Error bars indicate standard error that this effect was significant and not due to chance:

\[ t(463) = 4.4403; \ p < 0.05 \]

For experiment 2, shape error for the static viewing group averaged 125 pixels while the shape error for the gaze directed group averaged 85 pixels. The differences in the averages show that SGD w/shape modulation at the automatically generated targets
results in a lower shape error compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

\[ t(476) = 3.5317; p < 0.05 \]

Figure 4.9 shows how the shape error varies as the number of regions the participants were asked to recall increases for experiment 1. No clear trend is evident. Note, however, that the shape error for SGD w/shape modulation is consistently lower than that of static viewing. A similar effect was observed in experiment 2.

4.1.3.4 Percentage Gaze Time

Figure 4.10 shows the percentage of total gaze time spent within the target regions for the different groups of participants for each experiment.
In experiment 1, the static viewing group (control) spent 7.83% of the total gaze time within the relevant target regions while the gaze-directed group spent 12.34% (an increase of 57.6%). For comparison, the percentages from section 3.2 were 8.7% and 12.5% respectively (an increase of 43.7%). The differences in the percentages show that SGD w/shape modulation at pre-selected relevant regions results in more gaze time spent within the target regions compared to static viewing. This observation was expected and an independent-samples t-test confirms that the difference in gaze time in the target
regions was significant and not due to chance:

\[ t(498) = 8.5267; p < 0.05 \]

In experiment 2, the static viewing group spent 8.33% of the total gaze time within the relevant target regions while the gaze-directed group spent 14.29% (an increase of 71.54%). The larger percentage increase, compared to experiment 1, is related to the fact that there were fewer targets to consider in experiment 2. Given the fixed viewing time (10 seconds), it is easier to obtain wider coverage on a smaller number of targets. The differences in the percentages show that SGD w/shape modulation at the automatically generated targets results in more gaze time spent within the target regions compared to static viewing. An independent-samples t-test revealed that this effect was significant and not due to chance:

\[ t(384) = 8.8477; p < 0.05 \]

### 4.1.4 Conclusion

Actively guiding viewer attention to relevant information facilitates problem solving. Subtle gaze manipulation strategies are particularly useful as they do not require permanent or overt changes to the imagery in order to highlight the regions of interest. To help fully realize the benefits of gaze manipulation, this chapter in the dissertation addresses two important challenges, (1) the need for an automated approach for target selection in order to reduce manual intervention, and (2) the need for better gaze manipulation strategies to influence spatial learning.

We presented a novel framework that combines a saliency map with an edge map, and
image feature maps in order to facilitate both target prediction and better shape recollection. Our framework generates potential target regions and also provides a rough representation of the shape of the target object. We adapt the SGD technique to use this information to actively guide viewer attention during spatial information recollection tasks.

Our framework automatically predicted 81% of the target regions across 50 images that were task relevant. This is quite remarkable as the algorithm has no prior information of the task being assigned to the viewer. We have observed that SGD with shape modulation significantly improves accuracy of target count recall, spatial location recall, as well as shape recall. This is true whether only relevant targets were manually selected or a set of targets (containing approximately 20% false positives) are automatically generated using our framework.

The idea of predicting visual attention from scene context and task at hand further motivated a framework to solve the task inference problem. There is a need for a framework which can accurately classify the real-world task performed by the user using scene information and eye movement data.
Chapter 5

Real-World Task Inference using Eye Tracking

Humans perform numerous day-to-day tasks with ease. These real-world tasks are complex and involve a sequence of multiple sub-tasks for successful completion. The simple task of making a peanut-butter sandwich involves recognizing and interacting with multiple objects in our environment while planning each action in a specified order. It has been shown that eye movements are planned in advance which leads to motor actions for successful completion of the task at hand [159]. Humans look for things that are relevant for the current task and pay less attention to irrelevant objects in the scene. Researchers have shown that there is a high correlation between visual cognition and eye movements when dealing with complex tasks [141].

Tasks at hand have been shown to influence visual attention greatly. Yarbus [5] showed that eye movement not only depends on the scene presented but also on the viewer’s intent. Scene context also plays a vital role in attracting visual attention to specific regions in the viewer’s environment. The scene, and the observer’s eye movements,
provide vital information about the task being performed. Numerous studies have been conducted to understand how salient features and contextual information in a scene drive visual attention. The task inference problem in the context of this section can be stated as, *identify the task performed by a user using key features from the scene and a subject’s real-time eye movement data*. In this section, we attempt to combine features from the scene camera video and eye movement data to infer the task performed by the viewer.

### 5.1 Real-World Task Inference using Visual Bag-Of-Features and Eye-Movement Data

The ability to analyze and comprehend the task being performed by a user is very natural to us. The scene where the task is being performed, the actions of the user and their eye movements play a vital role in classifying the task at hand. Thorpe et al. [160] found that humans are highly capable of categorizing complex natural scenes that contain animals or vehicles in them. It has also been shown that humans require little or no attention to rapidly categorize a natural scene [161]. Several algorithms have been proposed to use computer vision techniques to automatically classify the scene from images and videos.

A technique to learn and categorize natural scene was developed using Bayesian hierarchical model that does not require expert annotations to train the system [162]. Other techniques use natural image statistics with context (low-level features) in learning scene categorization [163]. A widely used approach for scene classification is extracting salient image patches and constructing an image vocabulary (bag of visual words) [164]. The dimension of the patches, region in the image and weighting of the visual words play a
Figure 5.1: Figure shows the overall system design for task inference. Subjects’ gaze is tracked while performing real-world tasks. Visual bag-of-features and eye-movement features are extracted using the scene camera video from the eye-tracker. An SVM classifier is used to construct the corresponding trainer for task inference. The bag-of-features, eye-movement features, and eye-movement features + bag-of-features classifier is evaluated on subject test data videos.

Feature extraction plays a significant role in scene classification. A spatial pyramid matching technique was proposed for classifying 15 natural scenes using scale-invariant feature transforms (SIFT) [165]. This technique was more robust and also provided accurate natural scene categorization.

Feature extraction plays a significant role in scene classification. Feature extraction techniques find interesting points and important objects in the image. Feature extraction algorithms that are invariant to image translation, scaling, rotation, and partially invariant to illumination changes are highly preferred. Some of the feature detection algorithms are biologically inspired and share similar properties of object recognition in primate vision. Feature detection algorithms such as Scale-invariant feature transform
(SIFT) [156], Speeded-Up Robust Feature (SURF) [157], Maximally Stable Extremal Regions (MSER) [158], Histogram of Oriented Gradients (HOG) etc. are widely used for object detection and scene classification. The saliency of particular regions or objects in an image also provide key regions for classification. Many visual saliency models have been proposed to predict where humans focus their attention in an image [3, 152]. These salient regions have also been used for scene categorization and classification. In this chapter we classify the tasks performed by the viewer by building bag-of-features with spatial pyramid matching using SURF feature detection algorithm.

Along with image features we also use eye movement data for task inference. It is shown that the pattern of eye movements heavily depends on the task being assigned [6, 40, 166]. Eye movements have been used to study natural behavior [141], comprehension process [167], everyday activities [159] and usability [168]. Most models use eye movements for very specific task (on images) within a controlled laboratory environment [169, 170]. We present a model for task inference in a natural environment on real-world tasks using visual bag-of-features and task-based eye movements.

5.1.1 System Design

Figure 5.1 shows the system design for our real-world task inference model using visual bag-of-features and eye movement data. The system uses a head-mounted eye tracking device that captures the video of the scene being viewed and the eye movements of the human subject. The first stage in the process of task inference is to extract the visual bag-of-features from the scene.
5.1.1.1 Visual Bag-Of-Features Extraction

A visual bag-of-features is created from image patches which are extracted from the key fixation frames of the scene camera video. For each task-based scene video we collect image patches at the beginning, middle, and end of every fixation made by the viewer. The patches are extracted by taking a 100x100 pixel image region around the center of the fixation point mapped onto the scene-camera video frame.

A patch size of 100x100 pixels was chosen after a pilot study to see how much visual information could be captured depending upon varying patch sizes. If the patches were too big they become non-discriminatory and too generalized per task, but if they’re too small there are not enough discernible features in them. As a result, the selected patch size of 100x100 proved to be the best balance between the two. Another issue with bigger patch size is the time to process the bag-of-features. Processing the entire frame (960x720p) was computationally expensive as the bag-of-words extraction took nearly 17 hours.

Once all patches are collected for every subject and for each task, we then create a visual bag-of-features using the SURF algorithm. A spatial pyramid match kernel allows to precisely match two collections of features in a high-dimensional space. Features are extracted from each image at the same spatial scale but at three levels as shown in figure 5.2. We use the 3 levels and 400 channels for the resulting feature vector as presented by Lazebnik et al. [165]. The extracted feature vectors were trained using a support-vector-machine (SVM). The original data collected for all tasks were split into a training data-set (66.67%) and a testing data-set (33.33%). The training data-set was partitioned (3-fold) such that 65% of the data was used for training and remaining 35% was used for cross-validation. The output of the SVM is a trained classifier which can
Figure 5.2: SURF features are computed on each patch extracted from the key fixation frames of the scene camera video. An example of three-level pyramid is shows with three different features types. The image is subdivided at three different levels of resolution. We obtain the count of features in each spatial bin for each level of resolution and each channel in the image. Image reproduced from Lazebnik et al. [165].

be used to infer the task being performed by the user for the given scene video and eye movement data.

The visual bag-of-features technique is traditionally used in scene/object classification for natural images. In this section, we model a classifier to perform task inference using image patches extracted from a subject’s scene videos. Using the trained classifier, every new patch will be classified to one of the task categories. All the patches extracted from the videos help in accurate task inference based on binary classification statistics.
5.1.1.2 Eye-Movements Feature Extraction

Eye movements convey vital information of the cognitive process involved when performing day-to-day tasks. Eye movements are useful in revealing a shift in attention in a scene for a given task. A sequence of eye movements, therefore, highly relates to the task at hand that a user is partaking in. The difficulty and complexity of the task also significantly influences user eye movements. This is based on the assumption that eye movements and visual cognition are highly correlated [141]. Eye movement can be used as either data to understand the underlying cognitive process, or as a combination of features to infer user’s visual intent or action. Fixations and saccades are computed from the subject’s raw eye-movement data for each task. Key eye-movement features are extracted from the fixation-saccade data. In this section, we extract 10 key features described below for task inference:

- **Mean of Fixation Duration**: The average of all the user fixations for each task (in milliseconds).

- **Standard Deviation of Fixation Duration**: The standard deviation (SD) of all the user fixations for each task (in milliseconds).

- **Normalized Fixation Count**: The count of all the user fixations for each task, normalized over the task duration (in countseconds).

- **Mean of Saccade Magnitudes**: The average of all the user saccade magnitudes for each task (in pixels). Saccade magnitude is the distance in pixels between the centers of consecutive fixations.

- **Standard Deviation of Saccade Magnitudes**: The standard deviation (SD) of all the user saccade magnitudes for each task (in pixels).
• **Mean of Saccade Orientations**: The average of all the user saccade orientations for each task (in degrees). Saccade orientation is the angle made by the saccade vector. The saccade orientation is always between $0^\circ$ and $90^\circ$.

• **Standard Deviation of Saccade Orientations**: The standard deviation (SD) of all the user saccade orientations for each task (in degrees).

• **Normalized Horizontal Saccade Count**: The count of all the user saccades with an orientation $\geq 0^\circ$ and $< 30^\circ$ normalized over the task duration (in $\text{countseconds}$).

• **Normalized Diagonal Saccade Count**: The count of all the user saccades with an orientation $\geq 30^\circ$ and $< 60^\circ$ normalized over the task duration (in $\text{countseconds}$).

• **Normalized Vertical Saccade Count**: The count of all the user saccades with an orientation $\geq 60^\circ$ and $\leq 90^\circ$ normalized over the task duration (in $\text{countseconds}$).

The horizontal and vertical saccades provide additional robustness to help differentiate tasks that involve a greater number of horizontal eye movements (watching a tennis match) vs. vertical eye movements (watching a ball bounce).

### 5.1.1.3 Combining Visual Bag-Of-Features and Eye-Movement Features

Humans effectively perform the task at hand by capturing scene information (need-based) and then make task-relevant eye movements. To accurately infer the task being performed by the subject, it is essential that we combine the visual bag-of-features and the eye-movement features together. The eye-movement features are appended to the
bag-of-features and the resultant feature vector is trained using SVM as discussed above. We compute the confusion matrix and perform binary classification statistics on the test data to evaluate the combined visual bag-of-features and eye-movement features as opposed to only using either bag-of-features or eye-movement features. We hypothesize that using a combined bag-of-features and eye-movement features approach will yield better accuracy for real-world task inference.

5.1.2 Experiment Design

In this section of the dissertation, we describe the experiment design for real-world task inference using bag-of-features and eye movement features.

5.1.2.1 Participants

12 subjects (9 males, 3 females), between the age group of 18 and 34 (avg. 24) volunteered to participate in the study. All participants reported normal or corrected-to-normal vision with no color vision abnormalities. None of the participants chosen in this study wore glasses (2 subjects had contact lenses). 6 participants (4 males, 2 females), were chose from West Virginia University and 6 participants (5 males, 1 female) were chosen from Oregon State University. The experiment was conducted in two different universities to ensure that the training and testing data did not come from the same environment setup.

5.1.2.2 Tasks

Figure 5.3 shows the 6 different real-world tasks performed by the participants while being eye tracked:
Figure 5.3: The 6 real-world tasks performed by participants from university A. The tasks are Counting Chairs, Having Conversation, Making Coffee, Walking in Corridor, Washing Hands and Writing Name

- **Counting Chairs**: Subjects were asked to count the number of chairs in the room. There were 5 chairs in the room at different orientations and distances from the subject.

- **Having Conversation**: Subjects were provide a topic and were asked to have a conversation with more than 2 people in the room. The subjects were either sitting or standing while having the conversation.

- **Making Coffee**: Subjects were asked to make coffee from the coffee machine. They were provided with brewed coffee, sugar, creamer, cup and stirrer. No specific order for the task of making coffee was provided.

- **Walking in Corridor**: Subject were asked to walk from one end of the building to the other through a corridor. Only the path was specified and no other instructions were given.
• **Washing Hands:** Subjects were asked to wash their hands. The wash room had a hand soap dispenser and paper towels to wipe their hands. No specific order for the task of washing hands was provided.

• **Writing Name:** Subjects were asked to write their name on the board in a classroom.

### 5.1.2.3 Procedure

After the participant volunteered for the study they were clearly briefed about how the head mounted eye tracker works and also the description of the 6 tasks to be performed. After the participant provided their consent for the study, the head mounted eye tracker was placed on the subject and their gaze was monitored in real-time. The eye-tracker used in this study is a SensoMotoric Instruments eye tracking glasses v2.0. The eye tracking glasses operate at sampling rate of 30Hz (Binocular). The manufacturer’s gaze tracking accuracy is 0.5° under normal conditions with 80° horizontal and 60° vertical gaze tracking range. The front-facing scene camera video was operated at 906x720p @ 30fps with high sensitivity to low light. The eye tracker was connected to a laptop powered by a dual core processor with 4GB RAM. The participants had the calibration procedure explained to them for the eye tracking glasses and a quick 3-point calibration was performed. Subjects were requested to fixate on dummy targets to validate their gaze tracking accuracy. After the calibration and validation results were satisfactory (< 1°) the participants were taken to the task area. The calibration and validation procedures were performed before each task. This was done to ensure accurate gaze tracking and no track loss occurred between tasks. The front-facing scene camera video and eye tracking data were recorded and the scene camera video and its corresponding gaze data was tagged to both a subject ID as well as the task performed for analysis.
<table>
<thead>
<tr>
<th>True positive (TP)</th>
<th>Correct patches identified as relevant to the task at hand.</th>
</tr>
</thead>
<tbody>
<tr>
<td>False positive (FP)</td>
<td>Incorrect patches identified as relevant to the task at hand.</td>
</tr>
<tr>
<td>True negative (TN)</td>
<td>Incorrect patches identified as irrelevant to the task at hand.</td>
</tr>
<tr>
<td>False negative (FN)</td>
<td>Correct patches identified as irrelevant to the task at hand.</td>
</tr>
</tbody>
</table>

Table 5.1: Mapping of trained feature outcomes on test data to true positives, true negatives, false positives, and false negatives.

5.1.2.4 Evaluation

The data obtained from the head-mounted eye tracker was used to compute the bag-of-features and eye-movement features as explained in section 5.1.1. We use binary classification statistics to evaluate the features that accurately infer the task being performed by the subject. We compute the sensitivity, specificity, and overall accuracy for each of the techniques proposed. To calculate these properties we categorize the test outcomes as true-positives, true-negatives, false-positives, and false-negatives. Table 5.1 shows the mapping of trained feature outcomes to these conditions.

For this study sensitivity refers to the ability of the trained feature to correctly identify the task relevant patches. Sensitivity is computed as follows:

\[
Sensitivity = \frac{(\#ofTP)}{(\#ofTP + \#ofFN)} \times 100
\]  

(5.1)
On the other hand, *Specificity* refers to the ability of the trained features to correctly identify patches that are task irrelevant. Specificity is defined as follows:

\[ \text{Specificity} = \frac{(\# \text{of TN})}{(\# \text{of TN} + \# \text{of FP})} \times 100 \]  

(5.2)

The sensitivity and specificity values can then be combined to produce a binary classification based measure of balanced accuracy as follows:

\[ \text{Balanced Accuracy} = \frac{(\text{Sensitivity} + \text{Specificity})}{2} \]  

(5.3)

### 5.1.3 Results and Discussion

Figure 5.4 shows a subset of patches (100x100 pixels) extracted from the key fixation frames of the scene camera video. The patches from 8 randomly chosen subjects (66.67%) are used as training data and the remaining 4 subjects’ data (33.33%) are used as testing data-set.

#### 5.1.3.1 Results From Bag-Of-Features

Image patches for building the classifier were extracted using the key fixation frames of the scene camera video from 8 subjects. A bag-of-features is created using SURF and spatial-pyramid matching as explained in section 5.1.1.1. A SVM classifier is constructed on these bag-of-features using 65% of the data-set for training and remaining 35% for validation. After an optimum number iterations the trained classifier is then evaluated on the test data patches extracted from 4 subjects chosen before.
Figure 5.4: A subset of training patches randomly chosen for each task from 8 subjects. Each patch (100x 100 pixels) is extracted from key fixation frames of the scene camera video.

Figure 5.5 shows the confusion matrix (CM) for the training data from 8 subjects. Each row represents the classification of patches for the specified task using bag-of-features. The last column in the CM indicates the hit-rate for each task mentioned in that row. The task of Washing Hands has a hit-rate less than 50% as indicated in red. The average hit-rate for all the tasks for training data-set is 73%.

Figure 5.6 shows the confusion matrix (CM) for the testing data from 4 subjects. The tasks Washing Hands and Making Coffee have hit-rate less than 50%. Even when the hit-rate is less than 50%, the number of patches classified correctly as task relevant is always higher than patches incorrect classified for every other task. The average hit-rate for all the tasks for testing data-set is 67%. 
Table 5.2 shows the binary classification results for the test data using bag-of-features method. The sensitivity value shown in the table is same as the hit-rate computed in the confusion matrix in figure 5.6. The specificity value shows true negative rate and provides the measure for type II error for the classifier. The balanced accuracy and the sensitivity values show that tasks *Making Coffee* and *Washing Hands* are lower compared to other tasks. The reason for the mismatch of patches can be attributed to a lot of hand images in both these tasks. We analyze eye movement features to see if these two tasks can be classified accurately. The average balanced accuracy for all the tasks for bag-of-features classifier on the training data-set is 78.14%.
Figure 5.6: Confusion matrix for the testing data-set bag-of-features classifier. Each row represents the classification of patches for the specified task. The green values in the last column indicates hit-rate > 50% and red indicates hit-rate otherwise.

Table 5.2: Table showing binary classification statistics (sensitivity, specificity and balanced accuracy) for each task using bag-of-features. The average balanced accuracy for all the tasks for bag-of-features classifier on the training data-set is 78.14%.
Chapter 5 - Real-World Task Inference using Eye Tracking

5.1.3.2 Results From Eye-Movement Features

Eye-movement data provides useful information about the task being performed. We extract eye-movement features as described in section 5.1.1. For training data we compute the eye-movement feature vectors for each task video from 8 subjects randomly selected in subsection 5.1.1.1. A SVM classifier is trained using 48 feature vectors (8 subjects, 6 tasks) with 65% of the data-set for training and remaining 35% for validation. After an optimum number iterations the trained classifier is then evaluated on the 24 test data videos (4 subjects, 6 tasks) chosen in subsection 5.1.1.1 for testing.

Figure 5.7 shows the confusion matrix for the training results of the classifier using eye-movement features. The average hit-rate for all the tasks for training data-set is 79%.
Figure 5.8 shows the confusion matrix for the test results using only eye-movement classifier. Eye-movements provide additional key features that help differentiate the tasks being performed by subjects. However we can see that the task for *Writing Name* is misclassified as *Making Coffee*. The average hit-rate for all the tasks for training data-set is 66%.

Figure 5.9 shows a subject’s scanpath video while performing the task of *Making Coffee* and *Washing Hands*. These two tasks had a lower classification accuracy computed using the bag-of-features approach. However eye-movements-features incorrectly classified the task of *Washing Hands* to *Writing Name*. Figure 5.10 shows the same subject’s scanpath for the task of *Counting Chairs* and *Walking in Corridor* which have a high classification accuracy. Visually the scanpath for these tasks are different and the eye-movement

---

### Figure 5.8: Confusion matrix for the testing data-set using eye-movement features classifier.

<table>
<thead>
<tr>
<th></th>
<th>Counting Chairs</th>
<th>Having Conversation</th>
<th>Making Coffee</th>
<th>Walking Corridor</th>
<th>Washing Hands</th>
<th>Writing Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting Chairs</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Having Conversation</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Making Coffee</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Walking Corridor</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Washing Hands</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Writing Name</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

The green values in last column indicates hit-rate > 50% and red indicates hit-rate otherwise.
Chapter 5 - Real-World Task Inference using Eye Tracking

Figure 5.9: Scanpath image from a subject while performing the task of Washing Hands (on left) and Making Coffee (on right). This image illustrates visually how the scanpath vary based on the task at hand. These tasks had a lower classification accuracy computed using the bag-of-features approach.

Figure 5.10: Scanpath image from a subject while performing the task of Counting Chairs (on left) and Writing Name (on right). These tasks have high classification computed using the eye-movement features approach.

features enable accurate task classification. However, eye movements by themselves are not sufficient to differentiate multiple (complex) tasks. In figure 5.11 we see that the eye movements for the Making Coffee task is similar to Writing Name task. These two tasks have a low classification hit-rate computed using only eye-movement features. Bag-of-features and eye-movements individually do not provide an accurate classification of tasks. Hence it is essential to combine eye-movement features and bag-of-features for an accurate task inference approach.

Table 5.3 shows the binary classification results for the test data using eye-movement
Chapter 5 - Real-World Task Inference using Eye Tracking

Figure 5.11: Scanpath image from a subject while performing the task of Making Coffee (on left) and Walking in Corridor (on right). These tasks have high classification computed using the eye-movement features approach.

<table>
<thead>
<tr>
<th>Task</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Balanced Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting Chairs</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Having Conversation</td>
<td>75</td>
<td>100</td>
<td>87.5</td>
</tr>
<tr>
<td>Making Coffee</td>
<td>75</td>
<td>81.25</td>
<td>78.13</td>
</tr>
<tr>
<td>Walking in Corridor</td>
<td>75</td>
<td>100</td>
<td>87.5</td>
</tr>
<tr>
<td>Washing Hands</td>
<td>50</td>
<td>82.35</td>
<td>66.18</td>
</tr>
<tr>
<td>Writing Name</td>
<td>50</td>
<td>88.23</td>
<td>56.18</td>
</tr>
</tbody>
</table>

Table 5.3: Table showing binary classification statistics (sensitivity, specificity and balanced accuracy) for each task using eye-movement features. The average balanced accuracy for all the tasks for bag-of-features classifier on the training data-set is 79.25%.

features method. The balanced accuracy and the sensitivity values show that tasks Washing Hands and Writing Name are lower compared to other tasks. The reason for the misclassification can be attributed to a similar scanpath between the two tasks. We analyze eye-movement features to see if these two tasks can be classified accurately. The average balanced accuracy for all the tasks for bag-of-features classifier on the training data-set is 79.25%.
5.1.4 Results From Eye-Movement Features + Bag-of-Features

To improve the sensitivity, specificity, and overall accuracy of task inference, we combine eye-movement features and the bag-of-feature patches from key fixation frames explained in section 5.1.1. For training data we compute the eye-movement feature vectors and bag-of-feature vectors for each task video from 8 subjects randomly selected in subsection 5.1.1.1. A SVM classifier is constructed on these bag-of-features using 65% of the data-set for training and remaining 35% for validation. After an optimum number iterations the trained classifier is then evaluated on the test data patches extracted from 4 subjects chosen before in subsection 5.1.1.1.

Figure 5.12 shows the confusion matrix of the training results from the classifier using the combined eye-movement features and bag-of-features. The average hit-rate for all the tasks for training data-set is 93%. By combining the eye-movement features with the bag-of-features we reduced the false-positive classifications for each task.

Figure 5.13 shows the confusion matrix (CM) for the testing data from 4 subjects chosen from subsection 5.1.1.1. By combining eye-movement features and bag-of-features the hit-rate for all the tasks in the test data-set is > 50%. Also, the average hit-rate for all the tasks for training data-set is 87% (67% for bag-of-features classifier and 66% for eye-movement features classifier).

Table 5.4 shows the binary classification results for the test data using the combined eye-movement features and bag-of-features method. The balanced accuracy and the sensitivity values shows an improvement for all the tasks over using only bag-of-features or eye-movement features. The average balanced accuracy for all the tasks for bag-of-features classifier on the training data-set is 92.25% (78.14% for bag-of-features classifier and 79.25% for eye-movement features classifier).
### Figure 5.12: Confusion matrix for the training data-set using eye-movement features + bag-of-features classifier. Each row represents the classification of combined eye movement feature vector and bag-of-features for the specified task. The green values in last column indicates hit-rate > 50% and red indicates hit-rate otherwise.

<table>
<thead>
<tr>
<th></th>
<th>Counting Chairs</th>
<th>Having Conversation</th>
<th>Making Coffee</th>
<th>Walking Corridor</th>
<th>Washing Hands</th>
<th>Writing Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting Chairs</td>
<td>54</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Having Conversation</td>
<td>0</td>
<td>921</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Making Coffee</td>
<td>0</td>
<td>0</td>
<td>214</td>
<td>0</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Walking Corridor</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>156</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Washing Hands</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>213</td>
<td>0</td>
</tr>
<tr>
<td>Writing Name</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>22</td>
<td>5</td>
<td>249</td>
</tr>
</tbody>
</table>

### Table 5.4: Table showing binary classification statistics (sensitivity, specificity and balanced accuracy) for each task using the combined eye-movement features and bag-of-features. The average balanced accuracy for all the tasks for bag-of-features classifier on the training data-set is 92.25%.

<table>
<thead>
<tr>
<th>Task</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Balanced Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting Chairs</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Having Conversation</td>
<td>90.86</td>
<td>100</td>
<td>95.42</td>
</tr>
<tr>
<td>Making Coffee</td>
<td>86.98</td>
<td>93.46</td>
<td>90.22</td>
</tr>
<tr>
<td>Walking in Corridor</td>
<td>97.29</td>
<td>100</td>
<td>98.65</td>
</tr>
<tr>
<td>Washing Hands</td>
<td>69.6</td>
<td>91.44</td>
<td>83.91</td>
</tr>
<tr>
<td>Writing Name</td>
<td>81.48</td>
<td>89.10</td>
<td>85.29</td>
</tr>
</tbody>
</table>
Chapter 5 - Real-World Task Inference using Eye Tracking

Figure 5.13: Confusion matrix for the testing data-set using eye-movement features + bag-of-features classifier. Each row represents the classification of the combined eye-movement features and bag-of-features for the specified task. The green values in last column indicates hit-rate > 50% and red indicates hit-rate otherwise.

<table>
<thead>
<tr>
<th></th>
<th>Counting Chairs</th>
<th>Having Conversation</th>
<th>Making Coffee</th>
<th>Walking Corridor</th>
<th>Washing Hands</th>
<th>Writing Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Counting Chairs</td>
<td>23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Having Conversation</td>
<td>0</td>
<td>308</td>
<td>31</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Making Coffee</td>
<td>0</td>
<td>0</td>
<td>147</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>Walking Corridor</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>108</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Washing Hands</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>158</td>
<td>69</td>
</tr>
<tr>
<td>Writing Name</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>10</td>
<td>132</td>
</tr>
</tbody>
</table>

The results above enable us to conclude that by combining the eye-movement features and the bag-of-features we improve the sensitivity, specificity and overall balanced accuracy for all the proposed. The results also indicate that the features from each classifier complement each other and compensates for any misclassification in the other classifier. We hypothesize that eye-movement features will be effective on complex tasks whereas bag-of-features will prove to be useful on simple sub-tasks that make up these complex tasks. Therefore, by combining the two classifiers we obtain an efficient and accurate solution to the real-world task inference problem.
5.1.5 Conclusion

This chapter presents a real-world task inference framework using visual bag-of-features (spatial-pyramid matching) and subject’s task-based eye-movement data. A bag-of-features classifier was modeled from image patches (100 x 100 pixels) extracted from key fixation frames on the scene camera video. Eye movements features were extracted from subjects’ eye-movement data while performing each task. The average accuracy of all the tasks using only bag-of-features was 78.14%. Another classifier modeled using only the eye-movement features had an overall accuracy of 79.25%. Finally, we combined the bag-of-features and eye-movement features to develop a robust and accurate task inference model, which provided an accuracy of 92.25%. We evaluated all 3 models using binary classification statistics and found that by combining bag-of-features and eye-movement features we obtain an overall increase in the sensitivity and specificity metrics. Our combined features approach enable us to develop a framework to classify complex real-world tasks.
Chapter 6

Conclusion and Future Work

This dissertation presented and discussed research on visual attention with specific focus on the use of subtle visual cues to guide a viewer gaze and the development of algorithms to predict the distribution of a viewer’s gaze about a scene. Specific contributions of this work include: a framework for gaze guidance to enable problem solving and spatial learning, a novel algorithm for task-based eye movement prediction, and a system for real-world task inference using eye tracking.

A gaze guidance approach is presented that combines eye tracking with subtle imagespace modulations to guide viewer gaze about a scene. Several experiments were conducted using this approach and results show that gaze guidance improves short-term spatial information recall, task sequencing, training, and password recollection.

A model of human visual attention prediction is presented that uses saliency maps, scene feature maps, and task-based eye movements to predict regions of interest was also developed to guide viewer’s attention to predicted regions of interest for improved task performance. The automatic target prediction framework enabled in choosing targets
that matched 84% to the targets chosen by expert for the specified task. Finally, a framework for inferring real-world tasks was developed using image features and eye movement data. We were able to successfully classify 6 different real-world tasks performed by the viewer.

Overall, this dissertation naturally leads to an overarching framework, that combines all three contributions to provide a continuous feedback system to improve performance on repeated visual search tasks. Section 6.1 details a list of possible future works using the three key contributions mentioned in this dissertation.

### 6.1 Future Work

Figure 6.1 shows an illustration of a real-world visual search problem of TSA agent searching for suspicious objects at airports. This task is critical and requires dedicated visual attention and hence leads to human fatigue. A potential future work of this dissertation is to combine the real-world task inference framework and customize it for each user to help model an automatic task-based region prediction system for detecting suspicious objects. We can then leverage the technique of active subtle gaze guidance to guide the user to these task-based regions for additional scrutiny. Such a system can provide additional suggestions to the experts and aid their overall task performance. The proposed system would behave as a continuous feedback loop and will be more accurate over time with expert suggestions to the model over the training period.

Below is a list of other possible future work from gaze guidance, task-based eye movement prediction, and real-world task inference contributions presented in this dissertation:
Figure 6.1: Figure shows an illustration of a real-world image search problem of TSA agents looking for suspicious items at airports. An interesting future system can leverage by combining user specific task inference, automatic target prediction, and subtle gaze guidance to aid and improve visual search task.

- **Training simulations and virtual environments:** Training simulations often involve complex, cluttered environments. Key to the success of these simulations is the ability to rapidly attend to targets and extract relevant information from the abundance of information available. Pairing SGD with simulations could facilitate faster responses to those spatial locations deemed more significant for status understanding. Gaze guidance approach will enhance task performance and training in virtual and augmented reality environments.

- **Education and learning:** Often the amount of information presented to students in educational settings, be it in the classroom on over distance education, is overwhelming. Using SGD to localize the important information first could help the student build a better cognitive map of the scene they are exploring, thus
opening potential for an enhanced learning experience. This could also benefit those that have difficulty paying attention during learning.

- **Disaster/first responder training:** First responders, through necessity, often operate in sub-optimal conditions, especially from a visual standpoint when the field of view is obscured by smoke, fire or power failure. Training with SGD could aid such personnel to rapidly build better spatial maps of the environment they are serving before they experience the actual physical environment. For example, by drawing visual attention to escape routes, exit doors and/or fire extinguishers, SGD can provide the most pertinent information first. The SGD paradigm could also be used for training occupants of buildings, such as elderly residents of nursing homes or patients in hospitals enabling them to build spatial maps of exit routes that could be used in the case of an emergency.

- **Gaming:** When environments change rapidly and successful game play depends on successful navigation, which in turn depends on constructing an accurate spatial map of the environment, SGD could be employed to help game players to rapidly build a spatial map prioritizing the most important game features including enemy or friendly targets, and portals to new levels.

- **Prioritized rendering:** Applications such as architectural walk-through afford the viewer a platform to view a highly realistic visual representation of an environment (which may or may not exist in the real world). Such walk-throughs come at a high computational cost. SGD could be used to determine those features of the environment deemed most salient to the correct spatial understanding and channel
computing resources into those areas first, resulting in prioritized rendering so that those areas in need of most refinement are finished before areas of the scene that may not be noticeable or relevant.

- **Deep-learning based Scanpath prediction:** Deep networks are machine learning models that are biologically inspired and can be trained to generate eye movements using convolutional neural networks. Also due to the temporal nature of eye movement data (scanpath), these networks can be recurrent in nature, thus enabling us to generate human-like scanpath on new unseen images.

- **Generating expert scanpath:** Gaze guidance has been shown to improve training when novices are guided along an expert scanpath. However, eye tracking experts across all images for a visual search task is impossible. An automatic expert scanpath generation model will enable the system to train on a selected viewer to generate expert-like scanpath on new unseen images.

- **Inferring sub-tasks from complex tasks:** An important goal of task inference is to infer sub-tasks that make an overall task. Expert scanpaths vary while performing a visual search task. It will be important to analyze how each expert performs a sub-task using a real-world system that can classify these sub-tasks to differentiate variations among experts.
Appendix A

Eye Tracking Datasets

Table A.1 lists publicly available eye tracking datasets with associated stimuli images. Some of these datasets were used for image data, algorithm development and comparison with the results from the experiments described in this dissertation.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>KTH [171]</td>
<td>31</td>
<td>99</td>
<td>1024 x 768</td>
<td>5</td>
<td>70</td>
<td>18</td>
<td>EyeLink I</td>
<td>-</td>
<td>Head mount</td>
</tr>
<tr>
<td>LIVE DOVES [172]</td>
<td>29</td>
<td>101</td>
<td>1024 x 768</td>
<td>5</td>
<td>134</td>
<td>21</td>
<td>Forward Tech Gen. V</td>
<td>200</td>
<td>Bite bar</td>
</tr>
<tr>
<td>MCGill ImgSal [173]</td>
<td>21</td>
<td>235</td>
<td>640 x 480</td>
<td>-</td>
<td>70</td>
<td>17</td>
<td>Tobii T60</td>
<td>60</td>
<td>-</td>
</tr>
</tbody>
</table>
### Appendix A - Eye Tracking Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Participants</th>
<th>Stimuli Duration</th>
<th>Stimulus Size</th>
<th>Session Duration</th>
<th>Equipment</th>
<th>Chin Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT Benchmark [174]</td>
<td>39 300</td>
<td>300</td>
<td>≈1024 x 768</td>
<td>3 61 19</td>
<td>ETL 400</td>
<td>Chin rest</td>
</tr>
<tr>
<td>MIT CSAIL [140]</td>
<td>15 1003</td>
<td>1003</td>
<td>≈1024 x 768</td>
<td>3 61 19</td>
<td>-</td>
<td>Chin rest</td>
</tr>
<tr>
<td>TUD Image 1 [175]</td>
<td>20 29</td>
<td>Varying</td>
<td>10 70 19</td>
<td>iView X</td>
<td>50</td>
<td>Chin rest</td>
</tr>
<tr>
<td>VAIQ [176]</td>
<td>15 42</td>
<td>Varying</td>
<td>12 60 19</td>
<td>EyeTech TM3</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Bibliography


[93] S. Andolina, A. Santangelo, M. Cannella, A. Gentile, F. Agnello, and B. Villa. Multimodal virtual navigation of a cultural heritage site: the medieval ceiling of
steri in palermo. In *Proceedings of the 2nd conference on Human System Inter-
978-1-4244-3959-1.

[94] Erich Bruns, Benjamin Brombach, Thomas Zeidler, and Oliver Bimber. Enabling
mobile phones to support large-scale museum guidance. *IEEE MultiMedia*, 14:16–
25, April 2007. ISSN 1070-986X.

March: mobile augmented reality for cultural heritage. In *Proceedings of the 17th
ACM international conference on Multimedia*, MM ’09, pages 1023–1024, New

[96] Shih-Chun Chou, Wen-Tai Hsieh, Fabien L. Gandon, and Norman M. Sadeh. Se-
mantic web technologies for context-aware museum tour guide applications. In
*Proceedings of the 19th International Conference on Advanced Information Net-
working and Applications - Volume 2*, AINA ’05, pages 709–714, Washington, DC,

Digital museums and diverse cultural knowledges: Moving past the traditional

[98] Silvia Tomasi Velli. *Le immagini e il tempo: Narrazione visiva, storia e allegoria

[99] Ann McNamara, Reynold Bailey, and Cindy Grimm. Improving search task per-
formance using subtle gaze direction. In *Proceedings of the 5th Symposium on
Applied Perception in Graphics and Visualization*, APGV ’08, pages 51–56, New


remote eye tracking of radiologists searching for breast cancer: a comparison.


_Acad Radiol_, 3(2):137–144, 1996. ISSN 1076-6332 (Print); 1076-6332 (Linking).


[132] Eduardo E. Veas, Erick Mendez, Steven K. Feiner, and Dieter Schmalstieg. Directing attention and influencing memory with visual saliency modulation. In


