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

Srinivas Sridharan*
Rochester Institute of Technology

Reynold Bailey
Rochester Institute of Technology

Abstract

Humans rely heavily on spatial information to perform everyday tasks. Developing good spatial understanding is highly dependent on how the viewer's attention is deployed to specific locations in a scene. Bailey et al. [2009] showed that it is possible to influence exactly where attention is allocated using a technique called Subtle Gaze Direction (SGD). The SGD approach combines eye tracking with subtle image-space modulations to guide viewer gaze about a scene. The modulations are presented to peripheral regions of the field of view, in order to attract the viewer's attention, but are terminated before the viewer can scrutinize them with their high acuity foveal vision. It was observed that subjects who were guided using SGD performed significantly better in recollecting the count and location of target objects, however no significant performance improvement was observed in identifying the shape of the target objects [Bailey et al. 2012]. Also, in previous studies involving SGD, the target locations were manually chosen by researchers. This paper addresses these two limitations. We present a novel technique for automatically selecting target regions using visual saliency and key features in the image. The shape recollection issue is solved by modulating a rough outline of the target object obtained using an edge map composed from a pyramid of low spatial frequency maps of the original image. Results from a user study show that the influence of this approach significantly improved accuracy of target count recollection, location recollection, as well as shape recollection without any manual intervention. Furthermore our technique correctly predicted 81% of the target regions without any prior knowledge of the recollection task being assigned to the viewer. This work has implications for a wide range of applications including spatial learning in virtual environments as well as image search applications, virtual training and perceptually based rendering.

CR Categories: I.3.3 [Computer Graphics]: Picture/Image Generation—Display Algorithms; I.4.7 [Image Processing and Computer Vision]: Feature Measurements—feature representation

Keywords: Gaze manipulation, short-term memory, eye-tracking, image saliency

1 Introduction and Background

The ability to guide a viewer's attention has important applications in training, spatial learning, problem solving, image search, data visualization, and advertising [Walther et al. 2005; Wang and Spelke 2002; Qvarfordt et al. 2010; McNamara et al. 2009; McNamara et al. 2012; Sridharan et al. 2012]. Computer-based strategies for guiding visual attention can be classified as either overt or subtle. Overt techniques typically involve permanent and highly salient changes to the imagery to highlight areas of interest. Examples of overt techniques include the depth-of-field effect from traditional photography which brings different areas of an image in or out of focus, and the use of arrows and other direct markings as is commonly done with medical imagery [DeCarlo and Santella 2002; Grant and Spivey 2003; Thomas and Lleras 2007; Wang and Spelke 2002; Groen and Noyes 2010]. Such markings obscure other (potentially important) regions of the image and alter the overall viewing experience. Subtle techniques, on the other hand rely on temporary or subdued changes to the imagery to guide visual attention. For example, contrast, color, or luminance can be adjusted to subtly increase the saliency of target regions and reduce the saliency of surrounding regions while maintaining a natural appearance in order to increase the likelihood of the target regions being attended to [Veas et al. 2011]. Brief modulations (motion cues) presented to the peripheral regions of the field of view can also be used to orient the viewer's attention. Studies on different cuing techniques have found a similar effect on target detection either using subtle or explicit cues [Lu et al. 2012; Lu et al. 2014].

In this paper we focus on subtle cues - specifically brief motion cues as described by Bailey et al. [2009]. Their technique is called Subtle Gaze Direction (SGD) and it uses real-time eye tracking to monitor where the high acuity foveal vision of the viewer is deployed on the image. If the viewer is not attending to the desired target, then brief luminance modulations are applied to the target in the periphery of the field of view. The onset of motion on an otherwise static retinal image attracts the viewer's attention and results in a saccade towards the target. To maintain subtlety, the modulations are terminated before the viewer can scrutinize them with their high acuity foveal vision. This approach to gaze manipulation has been shown to be fast and accurate: viewers typically attend to target regions within 0.5 seconds of the onset of the modulation and the resulting fixations are typically within a single perceptual span of the target.

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 [Bailey et al. 2012]. 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 the Bailey et al. [2012] study. 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

*e-mail:sxs9716@rit.edu

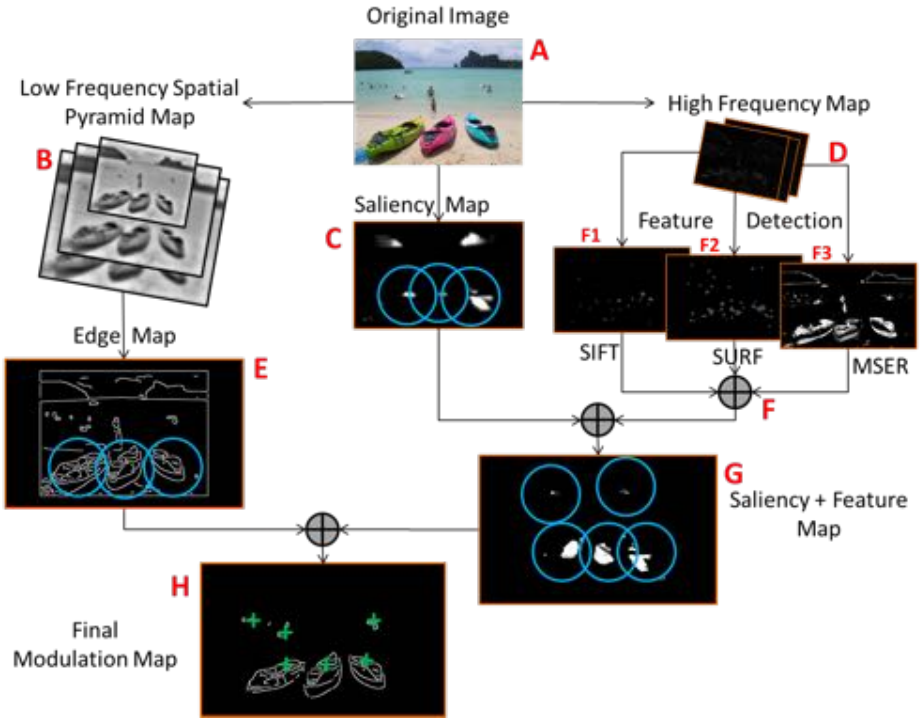

Figure 1: *Framework for automatically identifying regions of interest and shape boundaries for gaze guidance. An edge map, saliency map, and feature maps of the original image are combined to obtain the final modulation map. Source image courtesy of [knowphuket.com].*

approach, we observed a significant improvement in shape recollection as well as target count and location recollection.

In this paper, 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 preselected 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 *lableme* [Russell et al. 2008] which have been previously labeled by others, however significant manual effort is still necessary to build these databases. Other approaches rely on image segmentation [Felzenszwalb and Huttenlocher 2004; Garcia Ugarriza et al. 2009] or object detection [Viola and Jones 2001; Torralba et al. 2004] 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 [Itti and Koch 2000; Le Meur et al. 2007; Torralba et al. 2006; Achanta et al. 2009]. 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 our approach, 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 Bailey et al. [2012] study. 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.

The remainder of this paper is organized as follows: our automatic target generation framework is described in section 2, experiments to test the impact of our approach on spatial information recall are presented in section 3, analysis and discussion of the experimental results are presented in section 4, and the paper concludes in section 5 with a summary of the contributions and potential avenues of future research.

2 Automatic Target Prediction Framework

Our target prediction framework is illustrated in Figure 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:

- **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 [Canny 1986] 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 [2000] using normalized

center surround conspicuity maps obtained from image intensity, color (RGB) and Gabor orientated filters (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 [Lowe 2004], SURF [Bay et al. 2006] and MSER [Forssen and Lowe 2007] 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 complete list of parameters for the supporting computer vision algorithms is provided in the appendix.

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 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.

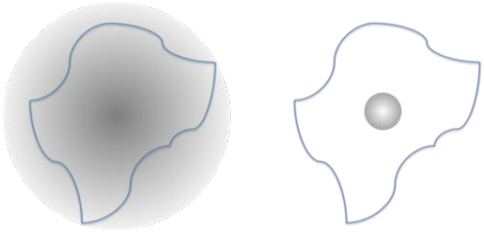

Figure 2: *Target modulation regions. Connected edge around the center is modulated using intensities computed from an enclosing circle of uniform Gaussian gradient (Left). Center of target is modulated using faint circular modulation with a uniform Gaussian distribution (Right).*

3 Experiment Design

We conducted two independent experiments using our adaptive SGD framework:

- **Experiment 1:** In experiment 1, we repeat the study conducted by Bailey et al. [2012] 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 paper.
- **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.

3.1 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 Bailey et al. [2012] study. 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 [Miller 1956] that the average human can only hold 7 ± 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 [Judd et al. 2012]. The number of objects or regions that the participants were asked to recall for each image ranged from 1 to 7.

3.2 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 modulation 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

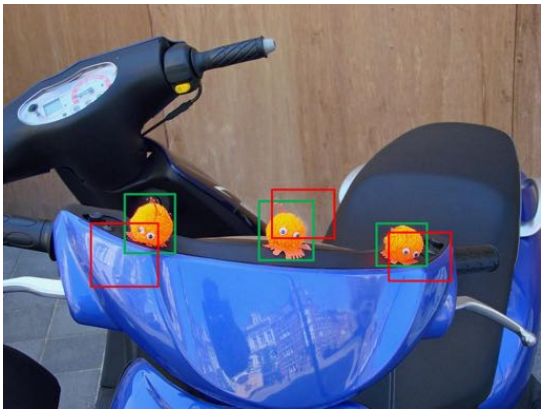

Figure 3: *One participant's solution (red) and correct solution (green) for an image from the training set. Participants were shown the image for 10 seconds then presented with a blank screen and asked to recall the location of each orange soft toy in the image by drawing rectangular regions on the blank screen. Image courtesy of [Judd et al. 2012]*

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 [McNamara et al. 2009].

3.3 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 taken in Bailey et al. [2012], 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 [Altmann 2004].

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).

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 $< 0.5^{\circ}$. 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 $\approx 65cm$ from the display. Modulations occurred at a rate of 20Hz. During modulation, the intensity of the modulated pixels varies along a sinusoidal curve between $\pm$ 10% of the original pixel's intensity value multiplied by the corresponding value from the relevant Gaussian distribution.

Bailey et al. [2012] defined three measures for spatial recall accuracy (counting error, location error, and shape error) that we also utilize in this paper. For convenience, the descriptions are repeated here:

"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 submit. Counting error is defined as follows:

$$|n - m| \tag{1}$$

- **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:

$$\frac{\sum_1^i \left(\sqrt{(x_{ai} - x_{bi})^2 + (y_{ai} - y_{bi})^2} \right)}{i} \tag{2}$$

  where i is the smaller of n and m and (x_{ai}, y_{ai}) and (x_{bi}, y_{bi}) represent the centroids of the ith 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:

$$\frac{\sum_1^i \left(|width_{ai} - width_{bi}| + |height_{ai} - height_{bi}| \right)}{i} \tag{3}$$

  where i is the smaller of n and m and $width_{ai}$ and $width_{bi}$ and $height_{ai}$ and $height_{bi}$ are the widths and heights of the ith closest pair of rectangles chosen from the set of actual solutions and participant solutions."

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

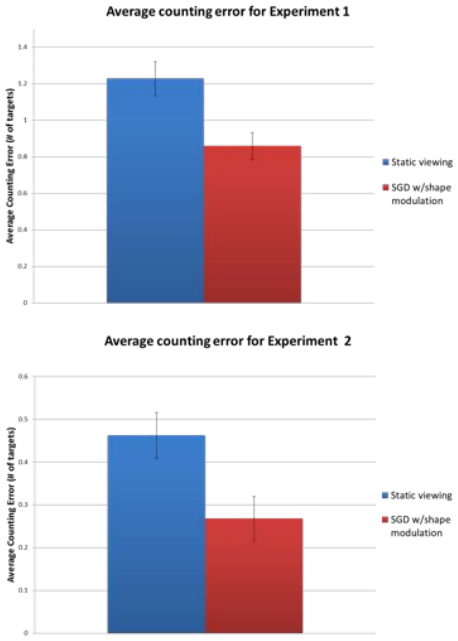

Figure 4: *Average counting error for participants from the static and gaze directed groups for experiment 1 (top) and experiment 2 (bottom). Error bars indicate standard error.*

4.1 Counting Error

Figure 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 Bailey et al. [2012] 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 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 Bailey et al. [2012]. A similar effect was also observed in experiment 2.

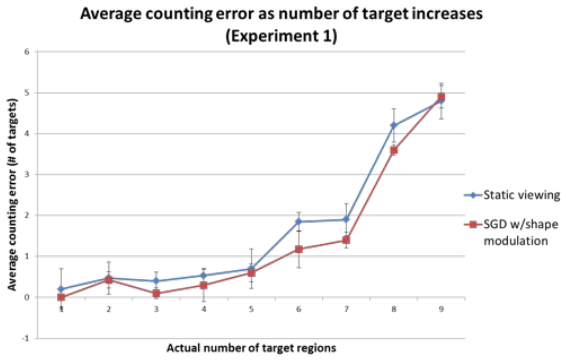

Figure 5: *Average counting error for participants from the static and gaze directed groups as number of target regions increase for experiment 1.*

4.2 Location Error

Figure 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 Bailey et al. [2012] were 134 pixels and 99 pixels respectively). The differences in the averages show that SGD w/shape modulation at preselected 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:

$$t(498) = 5.5145; p < 0.05$$

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 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.

4.3 Shape Error

Figure 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 Bailey et al. [2012] were 132 pixels and 120 pixels respectively). The differences in the averages show that SGD w/shape modulation at preselected relevant regions results in a lower shape error compared

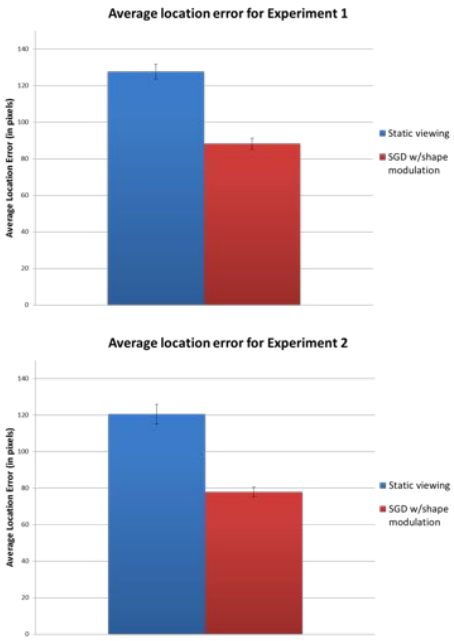

Figure 6: *Average location error for participants from the static and gaze directed groups for experiment 1 (top) and experiment 2 (bottom). Error bars indicate standard error.*

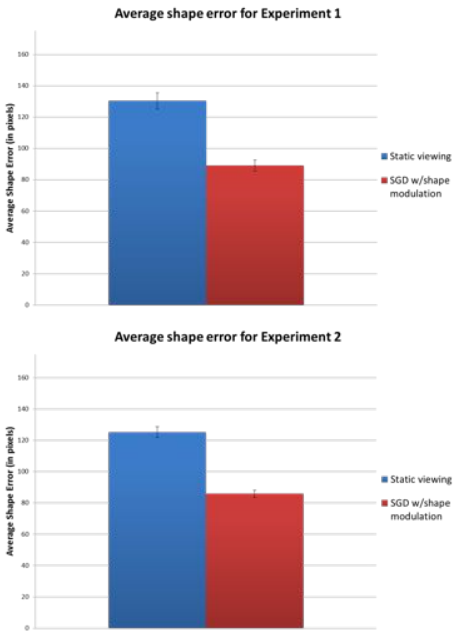

Figure 8: *Average shape error for participants from the static and gaze directed groups for experiment 1 (top) and experiment 2 (bottom). Error bars indicate standard error.*

to static viewing. An independent-samples t-test revealed 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 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.

4.4 Percentage Gaze Time

Figure 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 Bailey et al. [2012] were 8.7% and 12.5% respectively (an increase of 43.7%). The differences in the percent-

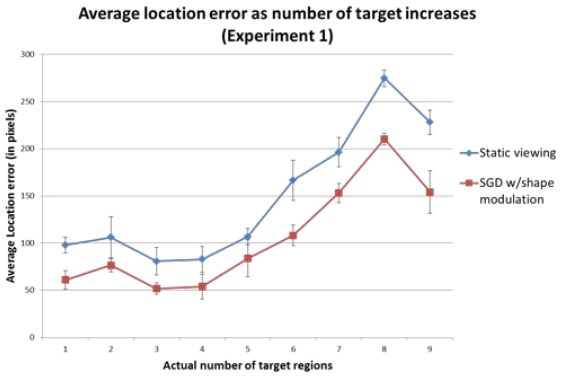

Figure 7: *Average location error for participants from the static and gaze directed groups as number of target regions increase for experiment 1.*

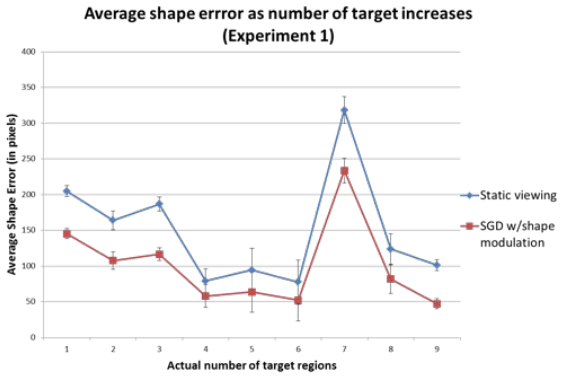

Figure 9: *Average shape error for participants from the static and gaze directed groups as number of target regions increase for experiment 1.*

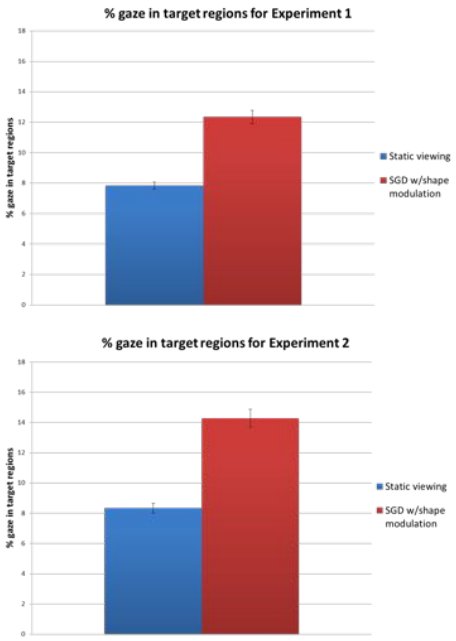

Figure 10: *Percentage gaze time spent within the target regions for participants from the static group and gaze-directed group with shape modulation for experiment 1.*

ages 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:

$$t(384) = 8.8477; p < 0.05$$

5 Conclusions and Future Work

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 paper 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 Subtle Gaze Direction (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 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.

An obvious next-step is to experiment with more complex and dynamic scenes. A key challenge in this regard is to enable our framework to perform in real-time. We also plan to explore ways to automatically prioritize the predicted target regions based on the stimuli image and the task at hand. Finally, We would also like to deploy our automatic target prediction and gaze-guidance framework in specialized visual search environments such as medical image analysis, baggage screening, and surveillance.

Acknowledgements

This material is based on work supported be the National Science Foundation under Award No. IIS-0952631. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Appendix: Algorithm Parameters

Below are the details for the computer vision algorithm parameters used in our automatic target prediction framework (see Figure 1):

- **Edge Map:** The canny edge detector was set with a threshold value of 0.4 and standard deviation of the Gaussian filter was 1.7 along both orientations.
- **Saliency Map:** We used the color, intensity and orientation maps with equal weights, a dyadic Gaussian pyramid and Gabor filter was implemented using 4 different orientations with a filter size of 9 pixels and standard deviation of 2. We also designed a winner take all network to obtain the final saliency map.
- **High Frquency Maps:** High pass filter with kernel sizes 7,15 and 31 are used to convolve the original image for feature map extraction.
- **SIFT:** The peak threshold value was set at 20 and the edge threshold was set to 7.5.
- **SURF:** The metric threshold was set to 800, with number of octaves as 5 and number of scales per octave to be 4.
- **MSER:** The step size between intensity threshold levels (Delta) was set to be 1.2, Size of the region in pixels were 30 to 14000 and the maximum area variation between extremal regions was set to 0.6.

References

ACHANTA, R., HEMAMI, S., ESTRADA, F., AND SUSSTRUNK, S. 2009. Frequency-tuned salient region detection. In *Com-*

puter Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 1597–1604.

ALTMANN, G. T. M. 2004. Language-mediated eye movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition 93*, B79–B87.

BAILEY, R., MCNAMARA, A., SUDARSANAM, N., AND GRIMM, C. 2009. Subtle gaze direction. *ACM Trans. Graph. 28* (September), 100:1–100:14.

BAILEY, R., MCNAMARA, A., COSTELLO, A., SRIDHARAN, S., AND GRIMM, C. 2012. Impact of subtle gaze direction on short-term spatial information recall. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, New York, NY, USA, ETRA '12, 67–74.

BAY, H., TUYTELAARS, T., AND VAN GOOL, L. 2006. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*. Springer, 404–417.

CANNY, J. 1986. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 6, 679–698.

DECARLO, D., AND SANTELLA, A. 2002. Stylization and abstraction of photographs. In *SIGGRAPH '02: Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, ACM Press, New York, NY, USA, 769–776.

FELZENSZWALB, P. F., AND HUTTENLOCHER, D. P. 2004. Efficient graph-based image segmentation. *International Journal of Computer Vision 59*, 2, 167–181.

FORSSEN, P.-E., AND LOWE, D. G. 2007. Shape descriptors for maximally stable extremal regions. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, IEEE, 1–8.

GARCIA UGARRIZA, L., SABER, E., VANTARAM, S. R., AMUSO, V., SHAW, M., AND BHASKAR, R. 2009. Automatic image segmentation by dynamic region growth and multiresolution merging. *Image Processing, IEEE Transactions on 18*, 10, 2275–2288.

GRANT, E., AND SPIVEY, M. J. 2003. Eye movements and problem solving: guiding attention guides thought. *Psychological Science 14*, 5, 462–466.

GROEN, M., AND NOYES, J. 2010. Solving problems: How can guidance concerning task-relevancy be provided? *Comput. Hum. Behav. 26* (November), 1318–1326.

ITTI, L., AND KOCH, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*, 10-12 (May), 1489–1506.

JUDD, T., DURAND, F., AND TORRALBA, A. 2012. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*.

KNOWPHUKET.COM. Digital image of sea kayaks. `http://www.knowphuket.com/islands.htm`.

LE MEUR, O., LE CALLET, P., AND BARBA, D. 2007. Predicting visual fixations on video based on low-level visual features. *Vision research 47*, 19, 2483–2498.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Intl. journal of computer vision 60*, 2, 91–110.

LU, W., DUH, B.-L., AND FEINER, S. 2012. Subtle cueing for visual search in augmented reality. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE Intl. Symposium on*, 161–166.

LU, W., DUH, H.-L., FEINER, S., AND ZHAO, Q. 2014. Attributes of subtle cues for facilitating visual search in augmented reality. *Visualization and Computer Graphics, IEEE Transactions on 20*, 3 (March), 404–412.

MCNAMARA, A., BAILEY, R., AND GRIMM, C. 2009. Search task performance using subtle gaze direction with the presence of distractions. *ACM Trans. Appl. Percept. 6* (September), 17:1–17:19.

MCNAMARA, A., BOOTH, T., SRIDHARAN, S., CAFFEY, S., GRIMM, C., AND BAILEY, R. 2012. Directing gaze in narrative art. In *Proceedings of the ACM Symposium on Applied Perception*, ACM, New York, NY, USA, SAP '12, 63–70.

MILLER, G. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review 63*, 81–97.

QVARFORDT, P., BIEHL, J. T., GOLOVCHINSKY, G., AND DUNNINGAN, T. 2010. Understanding the benefits of gaze enhanced visual search. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, ACM, New York, NY, USA, ETRA '10, 283–290.

RUSSELL, B. C., TORRALBA, A., MURPHY, K. P., AND FREEMAN, W. T. 2008. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision 77*, 1-3 (May), 157–173.

SRIDHARAN, S., BAILEY, R., MCNAMARA, A., AND GRIMM, C. 2012. Subtle gaze manipulation for improved mammography training. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ACM, New York, NY, USA, ETRA '12, 75–82.

THOMAS, L., AND LLERAS, A. 2007. Moving eyes and moving thought: on the spatial compatibility between eye movements and cognition. *Psychonomic bulletin and review 14*, 4, 663–668.

TORRALBA, A., MURPHY, K. P., AND FREEMAN, W. T. 2004. Sharing features: efficient boosting procedures for multiclass object detection. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, vol. 2, IEEE, II–762.

TORRALBA, A., OLIVA, A., CASTELHANO, M. S., AND HENDERSON, J. M. 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review 113*, 4 (Oct.), 766–786.

VEAS, E. E., MENDEZ, E., FEINER, S. K., AND SCHMALSTIEG, D. 2011. Directing attention and influencing memory with visual saliency modulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, CHI '11, 1471–1480.

VIOLA, P., AND JONES, M. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, IEEE, I–511.

WALTHER, D., RUTISHAUSER, U., KOCH, C., AND PERONA, P. 2005. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Comput. Vis. Image Underst. 100*, 1-2 (Oct.), 41–63.

WANG, R. F., AND SPELKE, E. S. 2002. Human spatial representation: insights from animals. *Trends in Cognitive Sciences 6*, 9, 376 – 382.