Eye and Gaze Tracking for Interactive Graphic Display

Qiang Ji
Dept. of Electrical, Computer, and Systems Eng.
Rensselaer Polytechnic Institute
qji@ecse.rpi.edu

Zhiwei Zhu
Dept. of Computer Science
Univ. of Nevada, Reno
zhu_z@cs.unr.edu

ABSTRACT

This paper describes preliminary results we have obtained in developing a computer vision system based on active IR illumination for real time gaze tracking for interactive graphic display. Unlike most of the existing gaze tracking techniques, which often require assuming a static head to work well and require a cumbersome calibration process for each person, our gaze tracker can perform robust and accurate gaze estimation without calibration and under rather significant head movement. This is made possible by a new gaze calibration procedure that identifies the mapping from pupil parameters to screen coordinates using the Generalized Regression Neural Networks (GRNN). With GRNN, the mapping does not have to be an analytical function and head movement is explicitly accounted for by the gaze mapping function. Furthermore, the mapping function can generalize to other individuals not used in the training. The effectiveness of our gaze tracker is demonstrated by preliminary experiments that involve gaze-contingent interactive graphic display.

Keywords

eye tracking, gaze estimation, HCI, Interactive Graphic Display

1. INTRODUCTION

Gaze determines the user's current line of sight or point of fixation. The fixation point is defined as the intersection of the line of sight with the surface of the object (such as the screen) being viewed. Gaze may be used to interpret the user's intention for non-command interactions and to enable (fixation dependent) accommodation and dynamic depth of focus. The potential benefits for incorporating eye movements into the interaction between humans and computers are numerous. For example, knowing the location of a user's gaze may help a computer to interpret a user's request and possibly enable a computer to ascertain some cognitive states of the user, such as confusion or fatigue.

In addition, real time monitoring of gaze position permits the introduction of display changes that are contingent on the spatial or temporal characteristics of eye movements. Such methodology is referred to as gaze contingent display paradigm. For example, gaze may be used to determine one's fixation on the screen, which can then used to infer what information the user is interested in. Appropriate actions can then be taken such as increasing the resolution or increasing the size of the region where the user fixates. Another example is to economize on bandwidth by putting high-resolution information only where the user is currently looking.

Gaze tracking is therefore important for HCI and intelligent graphics. Numerous techniques have been developed including some commercial eyes trackers. Video-based gaze estimation approaches can be partitioned into head-based approach, ocular-based approach, and the combined head and eye approach. The head based approach determines eye gaze based on the head orientation. In [9], a set of Gabor filters is applied locally to the image region which includes the face. This results in a feature vector to train a neural network to predict the two neck angles, pan and tilt, providing the desired information about head orientation. Gaze estimation by head orientation, however, only provides a global gaze since one's gaze can still vary considerably given the head orientation.

Ocular-based approach estimates gaze by establishing the relationship between gaze and the geometric properties of the iris or pupil within the eyes. Specifically, the iris-based gaze estimation approach computes gaze by determining the iris location or shape distortions from its image while pupil-based approach determines gaze based on the relative spatial positions between pupil and the glint (cornea reflection). Neural networks have been used in the past for this task [1, 11]. The idea is to extract a small window containing the eye and feed it, after proper intensity normalization, to a neural network. The output of the network determines the coordinates of the gaze.

So far, the most common approach for ocular-based gaze estimation is based on the relative position between pupil and the glint on the cornea of the eye [2, 4, 5, 8, 7, 3]. Assuming a static head, methods based on this idea use the glint as a reference point, thus the vector from the glint to the center of the pupil will describe the gaze direction. While contact-free and non-intrusive, these methods work well only

Int. Symp. on Smart Graphics, June 11-13, 2002, Hawthorne, NY, USA.

for a static head, which is a rather restrictive constraint on the part of the user. Even minor head movement can fail these techniques. This poses a significant hurdle to natural human computer interaction (HCI). Another serious problem with the existing eyes and gaze tracking systems is the need to perform a rather cumbersome calibration process for each individual. The latest research efforts are aimed at overcoming this limitation. Researchers from NTT in Japan proposed [8] a new gaze tracking technique based on modelling the eyeball. Their technique significantly simplifies the gaze calibration procedure, requiring only 2 points to perform the necessary calibration. The method, however, still requires relatively stationary head, and there exists difficulty in acquiring accurate geometric eyeball model for each subject. Researchers from IBM [7] is also studying the feasibility of completely eliminating the need of gaze calibration procedure by using two cameras and by exploiting the geometry of eyes and their images. Other recent efforts [12, 3] also focus on improving the eye tracking robustness under various lighting conditions.

In this paper, we propose to improve this approach so that gaze tracking can be accomplished robustly, accurately, naturally, and without the need of calibration.

2. EYE TRACKING

Gaze tracking starts with eye tracking. For eye tracking, we track pupils instead. We use infrared LEDs that operate at a power of 32mW in a wavelength band 40nm wide at a nominal wavelength of 880nm. As in Ji [6], we obtain a dark and a bright pupil image by illuminating the eyes with IR LEDs located off (the outer IR ring) and on the optical axis (the inner IR ring), respectively. To further improve the quality of the image and to minimize interference from light sources other than the IR illuminator, we use an optical band-pass filter which has a wavelength pass band only 10nm wide. The filter has increased the signal-to-noise ratio significantly, compared to the case without using the filter. Figure 1 illustrates the IR illuminator consisting of two concentric IR rings and the band-pass filter.

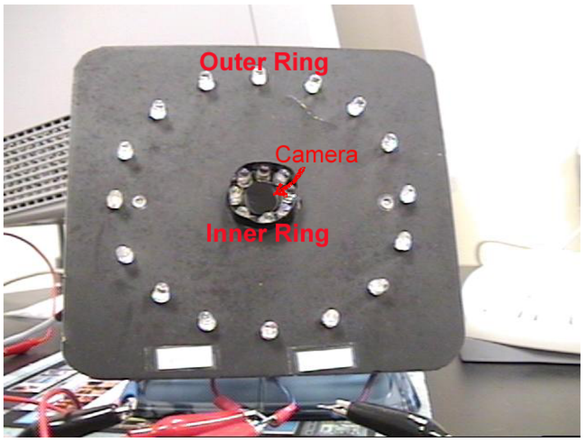


Figure 1: Hardware setup: the camera with an active IR illuminator.

Pupils detection and tracking start with pupils detection in the initial frames, followed by tracking. Pupil detection is accomplished based on both the intensity of the pupils (the bright and dark pupils as shown in Fig. 3 and on the appearance of the eyes using the support vector machine. The use of support vector machine (SVM) avoids falsely identifying a bright region as a pupil. Specifically, candidates of pupils are first detected from the difference image resulted from subtracting the dark pupil image from the bright pupil image. The pupil candidates are then validated using SVM to remove spurious pupil candidates. Given the detected pupils, pupils in the subsequent frames can be detected efficiently via tracking. Kalman filtering is used since it allows pupils positions in the previous frame to predict pupils position in current frame, therefore greatly limiting the search space. Kalman filtering tracking based on pupil intensity is therefore implemented. To avoid Kalman filtering go awry due to the use of only intensity, Kalman filtering is augmented by mean-shift tracking, which tracks an object based on its intensity distribution. Details on eye detection and tracking may be found in [12].

3. GAZE DETERMINATION AND TRACKING

Our gaze estimation algorithm consists of three parts: pupil-glint detection and tracking, gaze calibration, and gaze mapping. For this research, one's gaze is quantized into 8 regions on the screen (4 × 2) as shown in Fig.2.

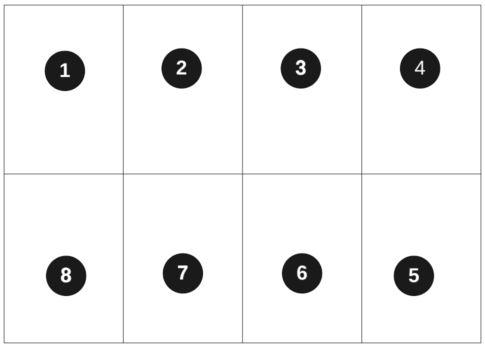
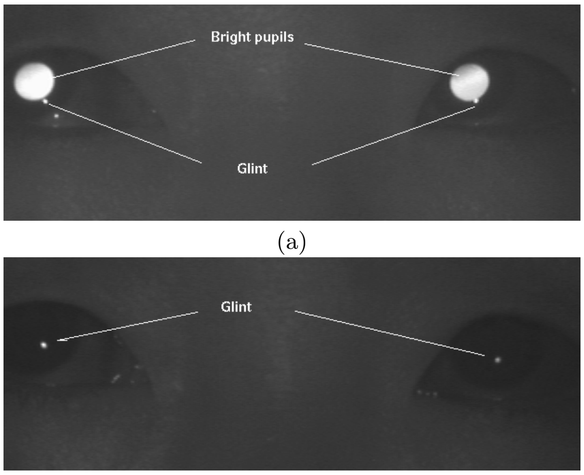


Figure 2: The quantized eye gaze regions on a computer screen

3.1 Pupil and Glint Detection and Tracking

Gaze estimation starts with pupil-glint detection and tracking. For gaze estimation, we continue using the IR illuminator as shown in Figure 1. To produce the desired pupil effects, the two rings are turned on and off alternately via the video decoder we developed to produce the so called bright and dark pupil effect as shown in Figure 3 (a) and (b).

Note glint (the small brightest spot) appears on both images. Given a bright pupil image, the pupil detection and tracking technique described in section 2 can be directly applied for pupil detection and tracking. The location of a pupil at each frame is characterized by its centroid. Algorithm-wise, glint can be detected much more easily from the dark image since both glint and pupil appear equally bright and sometimes overlap on the bright pupil image. In the dark image on the other hand, the glint is much brighter than the rest of the



(a)

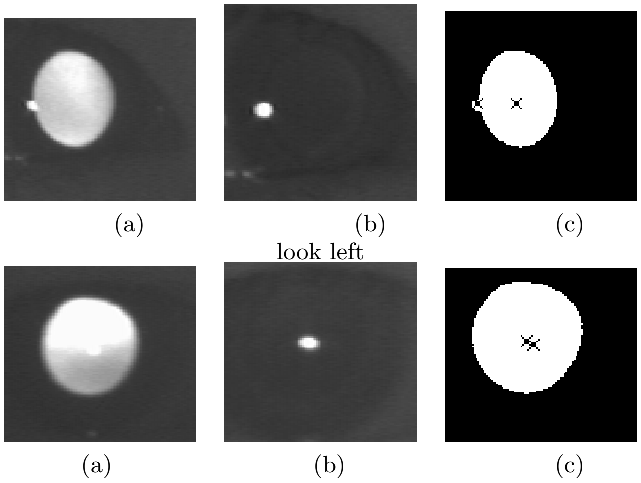
(b)

Figure 3: Bright (a) and dark (b) pupils images with glint.

eye image, which makes glint detection and tracking much easier. The pupil detection and tracking technique can be used to detect and track glint from the dark images. Figure 4 (c) show the detected glint and pupil.

3.2 Gaze Calibration

Given the detected glint and pupil, a mapping function is often used to map the pupil-glint vector to gaze (screen coordinates). Figure 4 shows the relationship between gaze and the relative position between the glint and the pupil. The mapping function is often determined via a calibration procedure. The calibration process determines the parameters for the mapping function given a set of pupil-glint vectors and the corresponding screen coordinates (gazes). The conventional approach for gaze calibration suffers from two shortcomings: 1) most of the mapping is assumed an analytical function of either linear or second order polynomial, which may not be reasonable due to perspective projection and spherical surface of the eye; 2) only coordinate displacements between pupil center and glint position are used for gaze estimation. This makes the calibration face orientation dependent. Another calibration is needed if the head has moved since last calibration, even for minor head movement. In practice, it is difficult to keep head still and the existing gaze tracking methods will produce incorrect result if the head moves, even slightly. Head movement must therefore be incorporated in the gaze estimation procedure. Another problem is that the mapping function derived for one person is not applicable to another. The calibration must, therefore, be performed for each individual. Here, we introduce a new gaze calibration based on Neural Network to overcome these two limitations.



(a) (b) (c)
look left

(a) (b) (c)
look frontal

Figure 4: Relative spatial relationship between glint and bright pupil center used to determine eye-gaze position. (a) bright pupil images, (b) glint images; (c) pupil-glint relationship generated by superimposing glint to the thresholded bright pupil images.

3.3 Gaze Calibration Via Generalized Regression Neural Networks (GRNN)

The reason to use NN is because of the difficulty in analytically deriving the mapping function that relates pupil and glint parameters to gaze under different face poses and for different persons. Given sufficient pupil and glint parameters, we believe there exists an unique function that relate gaze to different pupil and glint parameters.

Introduced in 1991 by D.L.Specht [10] as generalization of both radial basis function networks (RBFNs) and probabilistic neural networks (PNNs), GRNNs have been successfully used in function approximation applications. GRNNs are memory-based feed forward networks based on the estimation of probability density functions. GRNNs feature fast training times, can model non-linear functions, and have been shown to perform well in noisy environments given enough data. Our experiments with different types of NN also reveal superior perform of GRNN over the conventional feed forward back propagation NN.

The GRNN topology consists of 4 layers: the input layer, the hidden layer, the summation layer, and the output layer. The input layer has six inputs, representing the six parameters while the output layer has one node. The number of hidden nodes is equal to the number of training samples, with one hidden node added for each set of training sample. The number of nodes in the summation layer is equal to the number of output nodes plus 1. Figure 5 shows the GRNN architecture we use.

Due to significant difference in horizontal and vertical spatial gaze resolution, two identical GRNN networks were constructed, with output node representing the horizontal and vertical gaze coordinates s_x and s_y respectively.

The parameters to use for the input layer must vary with different face distances and orientations to the camera. Specifically, the input vector to the GRNN is

$$\mathbf{g} = \left[\begin{array}{cccccc} \Delta x & \Delta y & r & \theta & g_x & g_y \end{array}\right]$$

where Δx and Δy are the pupil-glint displacement, r is the ratio of the major to minor axes of the ellipse that fits to the pupil, θ is the ellipse orientation, and g_x and g_y are the glint image coordinates.

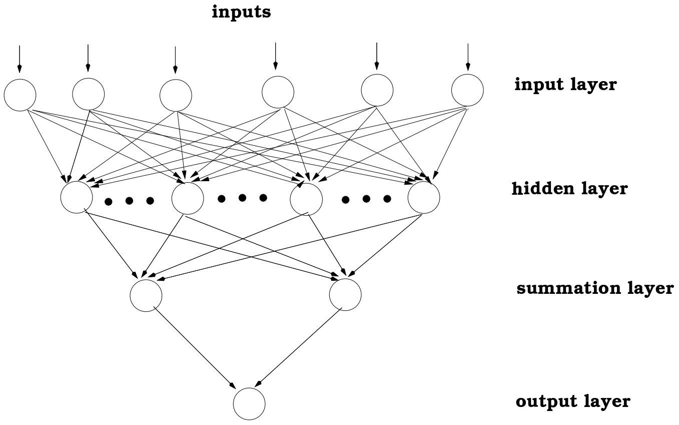


Figure 5: GRNN architecture used for gaze calibration

The choice of these input parameters is based on the following rational. Δx and Δy account for the relative movement between the glint and the pupil. The magnitude of the glint-pupil vector can also relate to the distance of the subject to the camera. r is used to account for face orientation. The ratio should be close to one when the face is frontal. The ratio becomes larger or less than 1 when the face turns either up/down or left/right. Angle θ is used to account for face rotation around the camera optical axis. Finally, (g_x, g_y) are used to account for the in-plane head translation. The use of these parameters allows to account for both head and pupil movement since head movement and pupil movement will introduce corresponding changes to these parameters. This effectively reduces the head movement influence. Furthermore, the input parameters are chosen such that they remain relatively constant for different people. For example, these parameters are independent of the size of the pupils, which often vary among people. The determined gaze mapping function, therefore, will be able to generalize. This effectively eliminates the need to re-calibrate for another person.

Before supplying to the GRNN, the input vector is normalized appropriately. The normalization ensures all input features are in the same range. A large amount of training data under different head positions are collected to train the GRNN. During the training data acquisition, the user is asked to fixate his/her gaze on each gaze region. On each fixation, 10 sets of input parameters are collected so that outliers can be identified subsequently. Furthermore, to collect representative data, we use one subject from each race including an Asian subject and a Caucasian subject. In the future, we will extend the training to additional races. The subjects ages range from 25 to 65. The acquired training data, after appropriate preprocessing (e.g., non-linear filtering to remove outliers) and normalization, is then used to train the NN to obtain the weights of the GRNN. GRNNs are trained using an one-pass learning algorithm and is therefore very fast.

3.4 Gaze Mapping

After training, given an input vector, the GRNN can then classify it into one of the 8 screen regions.

ground truth	estimated result (mapping target #)								Correct-ness
(target #)	1	2	3	4	5	6	7	8	rate (%)
1	94	6	0	0	0	0	0	0	94
2	2	90	8	0	0	0	0	0	90
3	0	3	88	7	0	2	0	0	88
4	0	0	3	96	1	0	0	0	96
5	0	0	0	0	96	4	0	0	96
6	0	0	1	0	7	90	2	0	90
7	0	0	0	0	0	5	89	6	89
8	0	0	0	0	0	0	2	98	98

Table 1: The classification results for gaze estimation with 100 testing gaze samples under different face poses for a person whose data is included in the training set.

ground truth	estimated result (mapping target #)								Correct-ness
(target #)	1	2	3	4	5	6	7	8	rate (%)
1	49	11	0	0	0	0	0	0	82
2	0	52	8	0	0	0	0	0	87
3	0	0	46	14	0	0	0	0	77
4	0	0	0	59	1	0	0	0	98
5	0	0	0	0	60	0	0	0	100
6	0	0	0	6	8	46	0	0	77
7	0	0	2	0	0	5	53	0	88
8	4	0	0	0	0	0	6	50	84

Table 2: The classification results for gaze estimation with 60 testing gaze samples for a new person whose data is not included in the training set.

4. EXPERIMENTAL RESULTS AND ANALYSIS

To validate the performance of our gaze tracker, we perform a series of experiments that involve the use of gaze to interactively determine what to display on the screen.

The first experiment involves visual evaluation of our eye tracking system. A laser pointer is used to point at the different regions of the computer screen. As expected, the user gaze is able to accurately follow the movement of the laser pointer which moves randomly from one gaze region to another gaze region, even under natural head movement.

To quantitatively characterize the accuracy of our system, the second experiment studies the performance of our system under different face orientations and distances to the cameras, and with different subjects. Tables 1 and 2 summarize the classification results. We can see our system can achieve an average of over 90% accuracy for the same subject under different face poses and an average of over 85% for a different subject.

Our study, however, shows that the system has some difficulty with older people, especially for those who suffer from some vision problem such as far-sighted or near sighted.

Our experiments show that our system, working in near real time (20 Hz) with an image resolution of 640 x480 on a Pentium 3, allows about 6 inches left/right and up/down head translational movement and allows ±20 degrees left/right head rotation as well as ±15 degrees up/down rotation. The distance to the camera ranges from 3.5 feet to 5 feet. The spatial gaze resolution is about 5 degrees horizontally and 8 degrees vertically, which correspond to about 4 inches horizontally and 5 inches vertically at a distance about 4 feet away from the screen.

Finally, we apply our gaze tracker to control graphic display on the screen interactively. This experiment involves the user gazes at a region of the computer screen and then blink 3 times, the region being gazed is then magnified to fill the screen. This repeats until the user can obtain enough details for the region of interest. One application this may be gaze-controlled map display as shown in Figures 6, 7, and 8, which show gaze-controlled map display at different levels of details. For a real-time demonstration of the gaze estimation, please refer to http://www.ecse.rpi.edu/homepages/qji/fatigue.html.

During study, we found that the vertical pupil movement range is much smaller than that of the horizontal range, causing the vertical glint-pupil vector measurement much more susceptible to external perturbation such as head movement. This leads to much lower SNR for the vertical data than that of the horizontal data, therefore leading to lower gaze vertical resolution. The current 4 x 2 gaze regions can be further refined to 4 x 3 or even 5 x 4. But this will lead to a decrease in tracking accuracy. This problem, however, can be overcome if we increase the image resolution.

5. CONCLUSIONS

In this paper, we present preliminary results for a new technique for gaze tracking. Compared with the existing gaze tracking methods, our method, though at a much lower spatial gaze resolution (about 5 degrees), has the following benefits: no calibration is necessary, allow natural head movement, and completely non-intrusive and obtrusive while still producing relatively robust and accurate gaze tracking. The improvement is resulted from using a new gaze calibration procedure based on GRNN. With GRNN, we do not need assume an analytical gaze mapping function and that we can account for head movements in the mapping.

While our gaze tracker may not as accurate as some commercial gaze tracker, it achieves sufficient accuracy even under large head movement and more importantly, it is calibration-free. It has significantly relaxed the constraints imposed by most existing commercial eye trackers. We believe, after further improvement, our system will find many applications including smart graphics, human computer interaction, and assistance of people with disability.

6. ACKNOWLEDGMENTS

The research described in this report is supported by a grant from AFOSR.

7. REFERENCES

[1] S. Baluja and D. Pomerleau. Non-intrusive gaze tracking using artificial neural networks. *Technical Report CMU-CS-94-102, Carnegie Mellon University*, 1994.

[2] Y. Ebisawa. Unconstrained pupil detection technique using two light sources and the image difference method. *Visualization and Intelligent Design in Engineering*, pages 79–89, 1989.

[3] Y. Ebisawa. Improved video-based eye-gaze detection method. *IEEE Transcations on Instrumentation and Measruement*, 47(2):948–955, 1998.

[4] T. E. Hutchinson. Eye movement detection with improved calibration and speed. *United States Patent [19]*, (4,950,069), 1988.

[5] T. E. Hutchinson, K. White, J. R. Worthy, N. Martin, C. Kelly, R. Lisa, , and A. Frey. Human-computer interaction using eye-gaze input. *IEEE Transaction on systems,man,and cybernetics*, 19(6):1527–1533, 1989.

[6] Q. Ji and X. Yang. Real time visual cues extraction for monitoring driver vigilance. *in ICVS 2001: Second International Workshop on Computer Vision Systems, Vancouver, Canada*, 2001.

[7] D. Koons and M. Flickner. Ibm blue eyes project. *http://www.almaden.ibm.com/cs/blueeyes.*

[8] T. Ohno, N. Mukawa, and A. Yoshikawa. Freegaze: A gaze tracking system for everyday gaze interaction. *Eye Tracking Research and Applications Symposium, 25-27 March, New Orleans, LA, USA*, 2002.

[9] R. Rae and H. Ritter. Recognition of human head orientation based on artificial neural networks. *IEEE Transactions on Neural Networks*, 9(2):257–265, 1998.

[10] D. F. Specht. A general regression neural network. *IEEE Transcations on Neural Networks*, 2:568–576, 1991.

[11] G. Yang and A. Waibel. A real-time face tracker. *Workshop on Applications of Computer Vision*, pages 142–147, 1996.

[12] Z. Zhu, K. Fujimura, and Q. Ji. Real-time eye detection and tracking under various light conditions. *Eye Tracking Research and Applications Symposium, 25-27 March, New Orleans, LA, USA*, 2002.

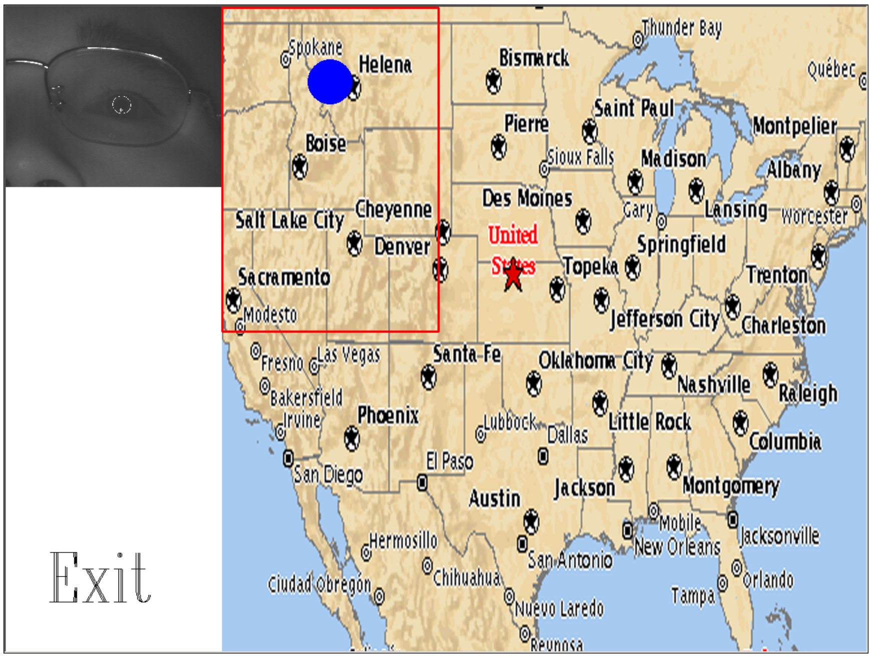


Figure 6: The map of the United States, with the gaze-selected region as marked by the shaded circle and the associated rectangle around it.

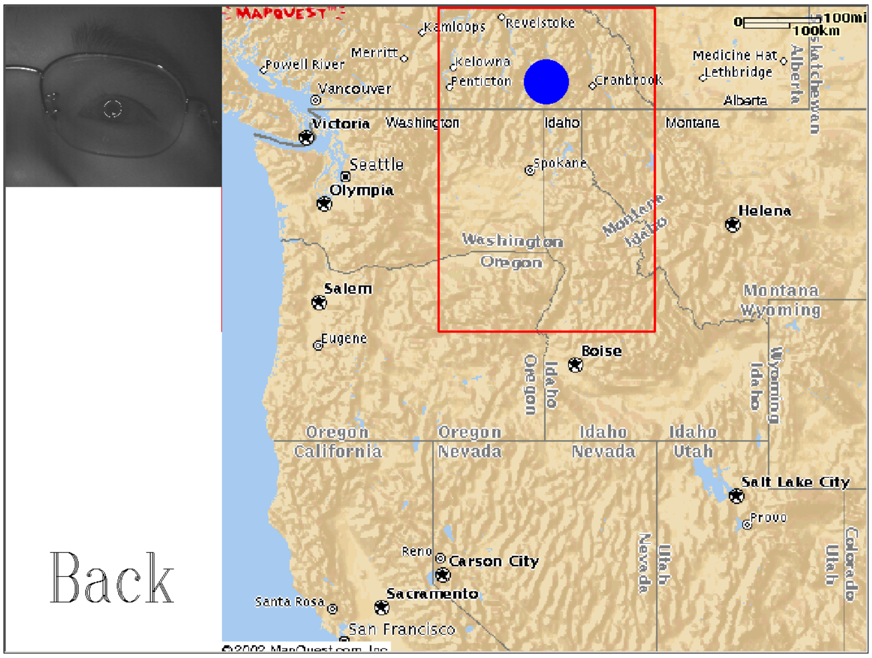


Figure 7: The blown-up area for the selected region in Fig. 6. Another selection is made by gaze on this image as indicated by the shaded circle and the associated rectangle around it.

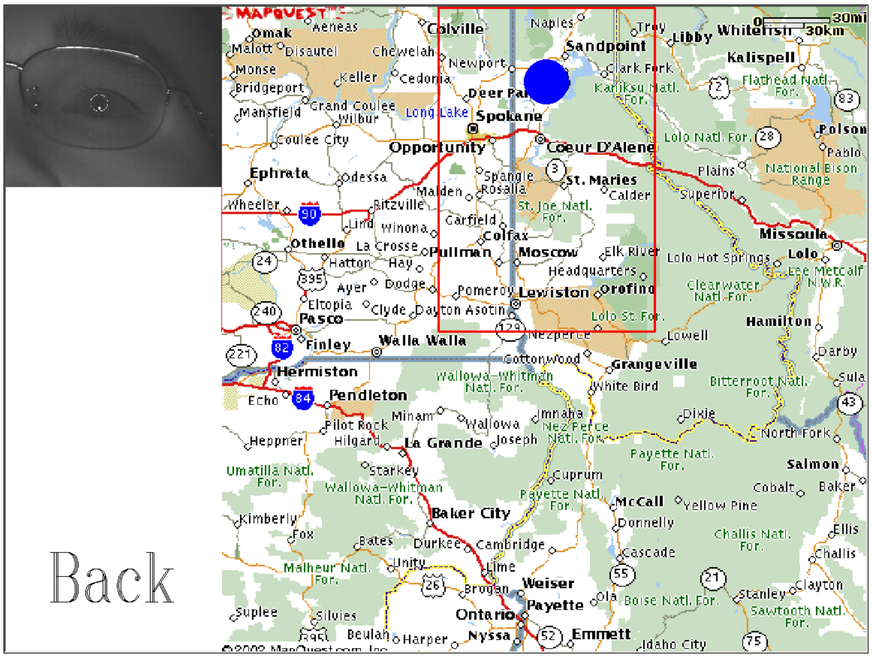


Figure 8: The blown-up area for the selected region in Fig. 7. Another selection is made by gaze on this image as indicated by the shaded circle and the associated rectangle around it.