A Teleoperation System Utilizing Saliency-Based Visual Attention

Wei-Chung Teng, Yi-Ching Kuo, Rayi Yanu Tara
Dept. of Computer Science & Information Engineering
National Taiwan University of Science and Technology
Taipei City, Taiwan
weichung@csie.ntust.edu.tw, rayi.yanu.tara@gmail.com

***Abstract*—We present a teleoperation system which supports foveated multi-resolution, wide field of view visual feedback. In the proposed system, a saliency-based visual attention system is implemented to simulate human visual system and to help determine the candidate visual attention area automatically. We link the visual attention system and the foveated multi-resolution image compression system to realize a real-time teleoperation system which requires less communication bandwidth to classic systems. Our experiment results show that the average operation time of the proposed system is only 57.9% comparing to a system with no visual attention clues in navigation tasks.**

Keywords-teleoperation system, visual attention, multi-resolution image, field of view

I. INTRODUCTION

A teleoperation system allows human operators to control remote robots, or teleoperators, to perform tasks in remote environments. Modern systems tend to rely on wireless communication channels to send back the information of remote sides. Although this approach provides high flexibility to the navigation ability of remote mobile robots, it suffers the problem of uneven signal strength and accordingly uneven bandwidth distribution in the navigation zone. On the other hand, the information that human operators most rely on among the different kinds of feedback is the visual feedback, yet unfortunately video stream requires most communication bandwidth. In fact, current video compression technology and wireless network standards can only sustain VGA resolution image with 40 degree or less field of view (FOV) [1]. Connectivity still remains as an open issue.

Since video feedback plays the most important role and requires the most bandwidth, it is essential to employ an effective compression method when the data is transmitted via network of limited and possibly unstable bandwidth. Besides the classic video compression methods like MPEG-4, another thread of researches utilizes the space variant resolution of retina. It is known that the resolution around fovea, the center of retina, is the highest, and it decreases progressively as the angle increases. Kortum and Geisler have indicated that compression methods utilizing this feature are much more efficient than traditional way [2]. In this paper, we adopt this approach to generate the image sequences. Since the outer part of an image needs much less space when compressed with foveated multi-resolution method, wide FOV visual feedback become affordable in our system. Finally, instead of asking the human operators to specify the region of interest, a saliency-based visual attention model is used to predict where the operators might put attention to in every frame. Therefore, the proposed teleoperation system is considered to help operators to manipulate the remote robots more efficiently.

II. BACKGROUND

A. Foveation Model

There are few measurement metrics of human vision like minimum separable acuity, minimum perceptible acuity, minimum perceptible acuity etc. If the vision of fovea is 1, then the resolution is reduced to 0.4 at 5 degrees away from the point of fixation, and again reduced to 0.2 at 10 degrees [3]. Shidoji[4] and Harashima[5] utilized this human visual characteristic in teleoperation systems and remote video communication. They used cameras of higher resolution to captures the area near the point of fixation, and covered other FOV with cameras of lower resolution. However the two-resolution images are still coarse.

B. Multi-resolution Pyramid System

Kortum and Geisler proposed a Foveated Imaging System (FIS) in 1996 [1], and developed a new system, Foveated Multi-resolution Pyramid (FMP) in 1998 [6]. Their method improves block artifacts and motion aliasing in the low resolution regions. FMP is a special coding system [7] for lower image bandwidth. According to the space variant resolution of retina, the resolution around the point of fixation is highest and it decreases progressively as the angle increases. Thus, it can achieve the target for bandwidth reduction.

C. Visual Attention Model

Visual search is the guidance of attention towards a target item. There are different types of looking behaviors including top-down, volitional attention application and bottom-up, involuntary attention capture [9]. Bottom-up attention is generated from a saliency map, the maximum of saliency map corresponds to the most salient stimulus to which the focus of attention should be directed next. The origin model of saliency-based visual attention is developed by Koch and Ullman in 1985 [12]. In visual search process, either attention or eye saccade are both related with salient stimulus [13].

This work was supported under the National Science Council Grants 97-2221-E-011-134.

978-1-4799-0652-9/13/$31.00 ©2013 IEEE

Since Koch and Ullman proposed a model of saliency-based visual attention, many researchers began to study this model [13]. This model calculates the saliency map in the input image and selects the most salient region. This visual attention model does not consider the time of fixation and the number of times of fixation; it only computes the visual features of input images, such as color, intensity, orientation and motion etc. as all features are combined into a saliency map. In 1998, Itti, Koch and Niebur proposed the bottom-up visual attention model [17]. They only concerned with the localization of the stimuli to be attended ('where'), not their identification ('what'). According to their model visual features are computed using linear filtering at eight spatial scales which are created using dyadic Gaussian pyramids [21]. Each feature is computed by "center-surround" operations for feature maps.

Finally, iLab C++ Neuromorphic Vision Toolkit (iNVT) is a comprehensive set of C++ classes for the development of neuromorphic models of vision provided by Itti's lab [22]. Given a static image, iNVT system predicts the point of the highest salience, and draws the focus of attention and the attention order. Furthermore, current version supports the motion feature maps and flicker feature map [19], by which we can detect the location of salience in continuous image sequences.

III. FOVEATION MODEL AND USER ATTENTION

A. Multiresolution-based Panoramic Visual System

We developed the Multiresolution-based Panoramic Visual System (MPVS) based on the idea of FMP. The system hardware includes Pioneer P3-DX mobile robot, Point Grey Ladybug2 spherical camera, high-end notebook computer as server and a computer as client side, as shown in Fig. 1.

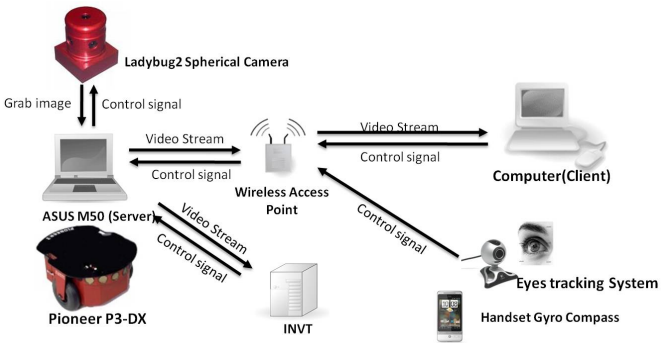

Figure 1. System configuration of the teleoperation system.

We set up a telescopic aluminium rack on Pioneer P3-Dx, as shown in Fig. 2, to simulate the user's point of view. The height of Ladybug2 camera is adjusted to 1.6 meter height. MPVS system uses OpenGL to stitch images from Ladybug2 and to generate a panoramic image of given FOV. It then produces a series of images of different resolutions by Mipmap function, and generates a multi-resolution image. Fig. 3 shows an example image generated by MPVS system. The red point represents the current location of fixation, default is at the bee. The spatial resolution declines dramatically and smoothly away from the bee.

Figure 2. The mobile robot with a spherical camera with 1.6 meters height.

Figure 3. An image generated by MPVS system.

B. How iNVT System Works with MPVS System

In the low-bandwidth limitation, we use a multi-resolution compression system (MPVS) to achieve image feedback in real-time, and apply visual attention model (iNVT) to the teleoperation system. iNVT can predict the location in the image which the operator will automatically and inconsciously direct his/her attention towards to. Therefore, the proposed teleoperation system can help operators to control the remote robots more efficiently then by keyboard.

The structure of these two systems is illustrated at Fig. 4. When the whole system starts, the multi-resolution image is generated by MPVS system will transmit to iNVT system. At the same time, MPVS system dealing with the latest image incessantly and sending images to user side until the location of salience is computed by iNVT system. After MPVS system receives the coordinate of salience form iNVT system, MPVS system processing the current image based on the corrdinate. Because of the image which is computed by iNVT system and the image which is processed by MPVS system is difference. We propose three methods to achieve real-time requirement: adjusting the computing time of iNVT system, reading one image limitation and inputting the lower resolution images.

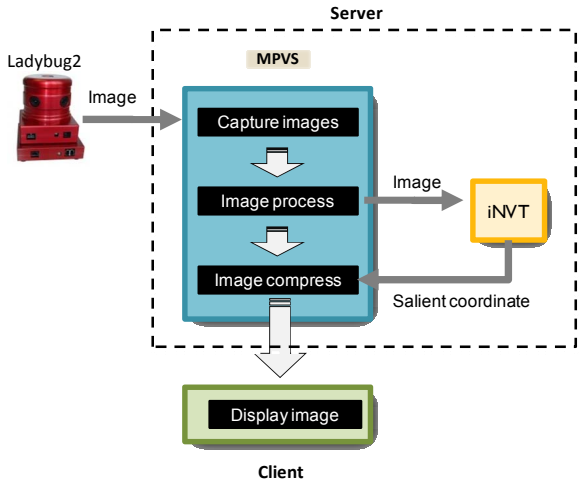

Figure 4. The mechanism of how iNVT system works with MPVS system.

C. *Adjusting the Computing Time of iNVT System*

iNVT system is not designed for real-time, therefore, to achieve our system in real-time, we adjusting the order of computing in iNVT system. Table I is the output of iNVT system before adjusting, and Table II is the output of iNVT system after adjusting. In Table I, we can detect two things. First, not every picture is generated salient coordinate by iNVT system. Second, the computing time is not fixed. In order to ensure our system in real-time, we limit the computing time, if the computing time is less than 100ms, then outputs the coordinate of salience, otherwise skips this image. As shown in Table II, each input frame can output its salient coordinate and the average computing time is less than 100ms.

TABLE I. THE OUTPUT OF iNVT SYSTEM BEFORE ADJUSTING

Frame number	Salient coordinate	Computing time (ms)
3	(169,91)	110.6
6	(335,212)	90.1
7	(210,96)	53.5
11	(331,67)	134.1
16	(108,52)	157.0
20	(411,42)	147.3
23	(310,151)	99.3
28	(457,175)	173.0
33	(493,43)	166.7
44	(366,160)	349.3

TABLE II. THE OUTPUT OF iNVT SYSTEM AFTER ADJUSTING

Frame number	Salient coordinate	Computing time (ms)
0	(325,111)	87.00
1	(325,111)	86.80
2	(323,109)	86.80
3	(323,109)	86.90
4	(321,107)	86.90
5	(321,107)	86.90
6	(321,107)	86.90
7	(319,106)	86.90
8	(319,106)	86.90
9	(319,106)	87.00

D. *Inputting the Lower Resolution Images*

The time an image is written into buffer is directly proportional to the image size. That is, bigger image size needs longer time to process. If we can reduce the image size, then it can cut down the image processing time. Therefore, we designed an experiment to observe the influence of the salient point on inputting the lower resolution image, and the result of the experiment is given by Fig. 5.

The left image in Fig. 5 is the original image, and the size is 680x680. The right image in Fig. 5 is the lower resolution image based on the left image. Furthermore, the salient location (yellow circle) of the two images seems the same. In fact, the coordinate of salience is a little difference between the two images, as shown in Table III. According to the experiment, we obtain a consequence: inputting the lower resolution images not only does not influence the coordinate of salience by iNVT system but also reduce the processing time.

Compare the delay time of our system between original image and the resolution of original image reduces to a factor of eight, and it is illustrated in Fig. 6. The blue curve stands for processing the original image and the red curve stands for processing the lower resolution image. And both of the curves stand for the difference in frames between the current frame and the frame is computed by iNVT system. Due to the producing time of a frame is approximately 100ms, the frame difference in the red curve is approximately one to two frames. That is, the real delay time is approximately 100ms to 200ms.

According to the results of our experiments, people hardly discover the 100ms to 200ms delay time of a frame. In addition, the average computing time of iNVT system largely decrease to 16.3ms. Hence, using above three methods can help our system to achieve real-time requirement.

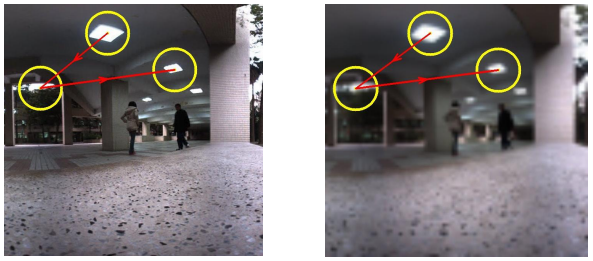

Figure 5. Compare the location of salience between original image (in left image) and the resolution of original image reduces to a factor of eight (in right image). Yellow circles represent the location of salience, and the red line represents the attention order.

TABLE III. THE COORDINATES OF SALIENCE IN FIG. 8.

Coordinates of left image	Coordinates of right image
(272,96)	(272,96)
(96,240)	(80,240)
(448,192)	(448,192)

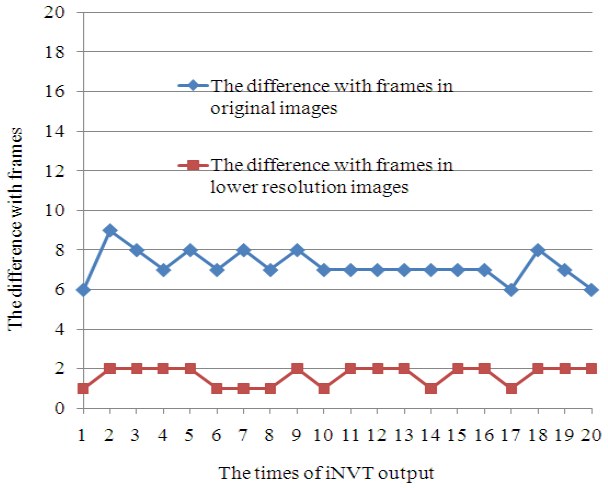

Figure 6. Compare the delay time of our system between original image and the resolution of original image reduces to a factor of eight. The blue curve stands for processing the original image and the red curve stands for processing the lower resolution image.

IV. Experiments and Discussion

A. Real-Time Experiment for iNVT System

The first experiment is designed to test if the iNVT system will take too much computing time and cause the visual feedback with attention mark not able to display in real-time base. The experiment subjects are talked to control the mobile robot to pass through a corridor as soon as possible. However, he/she must identify the text of colored papers pasted in walls. If iNVT system generates the salient location fast enough, subjects may have a change to look at the colored papers clearly while moving through the corridor. The text of colored papers as long as their locations changes randomly every time an experiment is conducted. Some scenes are shown in Fig. 7. According to the experiment results, the salient locations happen fast enough when the robot move no faster than 0.2 m/s.

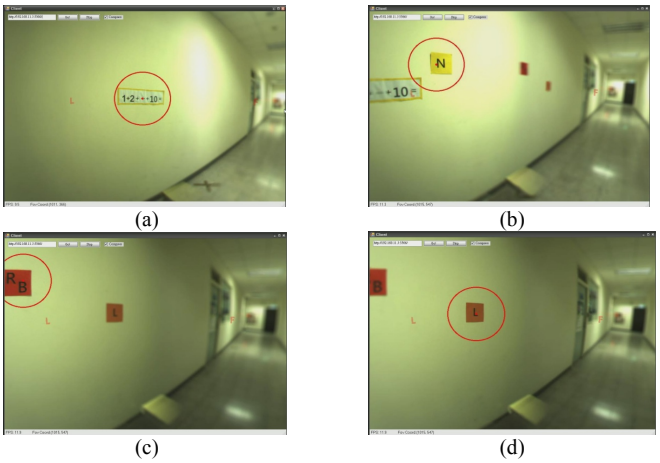

Figure 7. Two runs of the real-time experiment for iNVT system. Red circles represent the salient location. (a) and (b) belongs to the same run and (b) happens after (a). (c) and (d) belongs to the same run and (d) happens after (c).

B. Experiment of Operator Perception

To understand how the salient locations generated by iNVT system interest the operators, we designed two experiments for the operators control the Pioneer P3-DX to move through the corridor. The salient point of one experiment is set in the center of an image, and the salient point of the other experiment is generated by iNVT system. And the spatial resolution declines dramatically and smoothly away from the point of salience.

TABLE IV. A Questionnaire Is Designed to 30 Human Subjects

Does iNVT system help to understand the remote environment?
Are the gaze region provided by iNVT system interested to you?

TABLE V. The Score Range of Degrees

Degree	Score range
Very helpful	9 - 10
Helpful	7 - 8
Somewhat helpful	5 - 6
Little helpful	3 - 4
Not helpful	1 - 2

We designed a questionnaire as shown in Table IV. Feedback data from 30 subjects are collected as given by Table V. Furthermore, we gathered information from other 30 subjects by asking them to watch videos which are the results of the two aforementioned experiments. The collected results are given in Table VI. These 30 subjects deemed that iNVT system helps them understand the remote environment and the average score is 8.433. Moreover, the salient locations provided by iNVT system are interested to 30 human subjects with a average score of 8.166. Therefore, we conclude that most subjects agree that iNVT system is helpful to them.

TABLE VI. The Result from The Questionnaire

Questions	Score (Avg.)
Does iNVT system help to understand the remote environment?	8.433
Are the salient locations provided by iNVT system interested to you?	8.166
Remarks	30 subjects

C. Experiment of Operation Efficiency

We designed an experiment to understand if adding iNVT system to teleoperation systems whould help operators to control the remote robots in the keyboard more efficiently or not. The experiment environment is consists of several blue doorplates, several brown doors and a white wall. We paste up a white paper which is printed in a black AMO word on the wall, as shown in Fig. 8. And the users know that there has a paper which is printed in a black AMO word in the experiment environment, but they don't know the paper location, just control the robot to find the paper. Besides, the vision angle of

the robot is fixed, users can only control the robot to move forward.

Figure 8. The experiment environment consists of several blue doorplates, several brown doors and a white wall. And we paste up a white paper which is printed in a black AMO word on the wall.

There are two parts in the experiment of operation efficiency. One is the number of times users choosing the salient locations, and the other is the total operation time of users. In the number of times user choosing the salient locations, it differentiates between control group and experimental group. In control group, without iNVT system supported, the location of salience is chosen by the action of clicking the screen. And in the experimental group, with iNVT system supported, the location of salience also can be chosen by the action of clicking the screen. However, the location of salience can be determined by the action of users click the screen except iNVT system. Recording the click times of control group is illustrated in Fig. 9, and gathers the experimental data from ten subjects, as shown in Table X.

We analysed data in Table IX and the results are shown in Table VII. According to Table VII, we observed that the most reduction for original number of times is 0%, that is, the user does not determine the location of salience by clicking the screen in the experimental group, and the user finds the paper where we paste by iNVT system. And the least reduction for original number of times is 25%, and the count of explicitly assigning a new center of view is also declined to 16.9% in average comparing to the original counts.

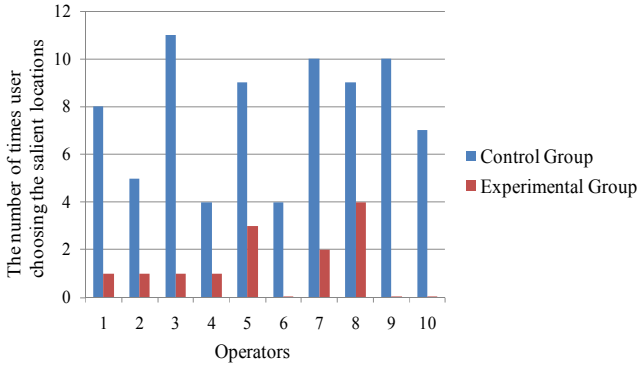

Figure 9. The results of the clicking times. There are ten subjects in this experiment as shown in x-axis, and the y-axis is the number of times user choosing the salient locations.

TABLE VII. ANALYSIS TABLE OF CLICKING TIMES

	Maximum	Minimum	Average
Reduce the number of times user choosing the gaze regions (%)	25%	0%	16.9%
Remarks	10 subjects		

Except the experiment of clicking times, we also design the experiment of total operation time. And it differentiates between control group and experimental group. In control group, without iNVT system supported, the location of salience is chosen by the action of clicking the screen. And in the experimental group, with iNVT system supported, the location of salience cannot be chosen by the action of clicking the screen. In other words, users totally depend on iNVT system to find the paper where we paste. Furthermore, recording the click times of control group is illustrated in Fig. 10, and gathers the experimental data from ten subjects, as shown in Table X.

We analyzed data in Table X and the results are shown in Table VIII. According to Table VIII, we can observe that the most operation time is reduced to 38.6% of original time, and the least operation time is reduced to 89.2% of original time, and the average operation time is reduced to 57.9% of original time. Table VII and Table VIII both show that applying iNVT system to the teleoperation system can reduce the number of times and the total operation time.

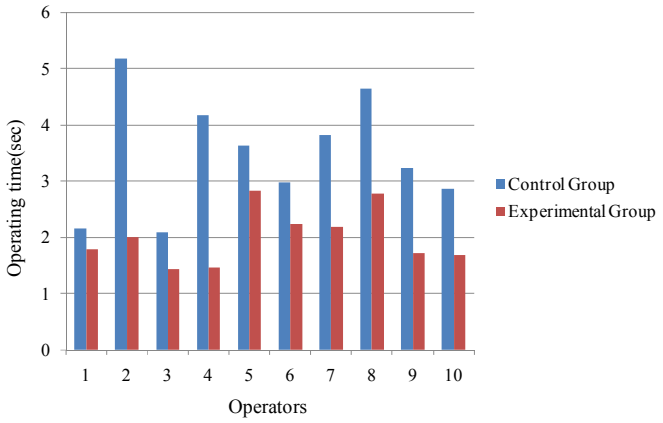

Figure 10. The results of the total operation time. There are ten subjects in this experiment, as shown in x-axis, and the y-axis is the total operation time until a subject finds the paper.

TABLE VIII. ANALYSIS TABLE OF TOTAL OPERATION TIME

	Maximum	Minimum	Average
Reduced operation time /original operation time (%)	89.2%	38.6%	57.9%
Remarks	10 subjects		

V. CONCLUSION

According to all our experiment results, the proposed teleoperation system can help operators to control the remote robots more efficiently. The average operating time is reduced to 57.9% of original time, and the count of explicitly assigning a new center of view is also declined to 16.9% in average comparing to the original counts. Besides, the visual attention model should be designed to prompts the subject where should be look in active form.

Future work includes adopting machine learning mechanism to our teleoperation system, and increasing the accuracy of the salient location. In addition, a top-down visual attention model in our teleoperation system will be developed in future, such as face detection. Besides, the proposed system may be applied on surveillance system. Since our system gives high salient values to human faces and moving objects (such as numbers on the license plate), users can monitor people or cars efficiently.

REFERENCES

[1] E. Guizzo, "iRobot and Cisco Team Up to Create Ava 500 Telepresence Robot," IEEE Spectrum, June 2013.

[2] P. Kortum, and W. Geisler, "Implementation of a foveated image coding system for image bandwidth reduction," Human Vision and Electronic Imaging, SPIE Proc., vol. 2657, pp. 350-360, 1996.

[3] T. Wertheim, "Über die indirekte sehschäfe," Zeitschrift für Psychologie und Physiologie der Sinnesorgane, vol. 7, pp. 172-178, 1894.

[4] T. Ienaga et al, "An effect of a large overlapped area of stereo pairs at the working point on a spatial multi-resolution stereoscopic video system," in Proc. of IEEE Conf. on Virtual Reality, pp. 277-278, 2005.

[5] T. Naemura, K. Sugita, T. Takano, and H. Harashima, "Multi-resolution stereoscopic immersive communication using a set of four cameras," Stereoscopic Displays and Virtual Reality Systems VII, Proc. of SPIE, vol. 3957, pp. 271-282, 2000.

[6] W. S. Geisler, and J. S. Perry, "A real-time foveated multiresolution system for low-bandwidth video communication," Human Vision and Electronic Imaging III, Proc. of SPIE, vol. 3299, pp. 294-305, 1998.

[7] Space Variant Imaging System <http://fi.cvis.psy.utexas.edu/>.

[8] P. Kortum, and W. Geisler, "Implementation of a foveated image coding system for image bandwidth reduction," Human Vision and Electronic Imaging, Proc. of SPIE, vol. 2657, pp. 350-360, 1996.

[9] J. M. Findlay, "Eye scanning and visual search," The Interface of Language, Vision, and Action: Eye Movements and the Visual World, New York: Psychology Press, 2004.

[10] J. M. Henderson, "Human gaze control during real-world scene perception," Trends in Cognitive Sciences, Elsevier, vol. 7, pp. 498-504, 2003.

[11] S. Treue, "Visual attention: the where, what, how and why of saliency," Current Opinion in Neurobiology, Elsevier, vol. 13, pp. 428-432, 2003.

[12] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," Human Neurobiological, 1985, vol. 4, pp. 219-227.

[13] L. Itti, C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," Vision Research, Elsevier, vol. 40, is. 10-12, pp. 1489-1506, 2000.

[14] L. Itti, C. Koch, "Computational modelling of visual attention," Nature Reviews Neuroscience, vol. 2, no. 3, pp. 194-203, March 2001.

[15] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," Vision Research, Elsevier, vol. 42, is. 1, pp. 107-123, 2002.

[16] R. J. Peters, A. Iyer, L. Itti and C. Koch, "Components of bottom-up gaze allocaiton in natural images," Vision Research, Elseview, vol. 45, is. 18, pp. 2397-2416, 2005.

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

[18] O. L. Meur, P. L. Callet, D. Barba and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28, is. 5, pp. 802-817, 2006.

[19] L. Itti, N. Dhavale, F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI, Proc. of SPIE, vol. 5200, pp. 64-78, 2004.

[20] D. Walther and C. Koch, "Modeling attention to salient proto-objects," Neural Networks, Elsevier, vol. 19, is. 9, pp. 1395-1407, 2006.

[21] H. Greenspan et al, "Overcomplete steerable pyramid filters and rotation invariance," in Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, pp. 222-228, June 1994.

[22] iLab neuromorphic vision c++ toolkit <http://ilab.usc.edu/toolkit/>.

TABLE IX. THE NUMBER OF TIMES USERS CHOOSING THE SALIENT LOCATIONS

Subjects	1	2	3	4	5	6	7	8	9	10
Control Group	8	5	11	4	9	4	10	9	10	7
Experimental Group	1	1	1	1	3	0	2	4	0	0

TABLE X. THE TOTAL OPERATION TIME OF USERS

Subjects	1	2	3	4	5	6	7	8	9	10
Control Group	02:09	05:11	02:05	04:10	03:38	02:59	03:49	04:39	03:14	02:52
Experimental Group	01:47	02:00	01:26	01:28	02:50	02:14	02:11	02:47	01:43	01:41