Scientific
Research
Publishing

Simulate Human Saccadic Scan-Paths in Target Searching

Lijuan Duan[1,2], Jun Miao[3]*, David M. W. Powers[4], Jili Gu[5], Laiyun Qing[6]

[1]Beijing Key Laboratory of Trusted Computing, Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data, College of Computer Science and Technology, Beijing University of Technology, Beijing, China
[2]National Engineering Laboratory for Critical Technologies of Information Security Classified Protection, Beijing, China
[3]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences(CAS), Institute of Computing Technology, CAS, Beijing, China
[4]School of Computer Science, Engineering & Maths, Flinders University of South Australia, Adelaide, South Australia
[5]Beijing Samsung Telecom R&D Center, Beijing, China
[6]University of Chinese Academy of Sciences, Beijing, China
Email: *jmiao@ict.ac.cn

Abstract

Human saccade is a dynamic process of information pursuit. There are many methods using either global context or local context cues to model human saccadic scan-paths. In contrast to them, this paper introduces a model for gaze movement control using both global and local cues. To test the performance of this model, an experiment is done to collect human eye movement data by using an SMI iVIEW X Hi-Speed eye tracker with a sampling rate of 1250 Hz. The experiment used a two-by-four mixed design with the location of the targets and the four initial positions. We compare the saccadic scan-paths generated by the proposed model against human eye movement data on a face benchmark dataset. Experimental results demonstrate that the simulated scan-paths by the proposed model are similar to human saccades in term of the fixation order, Hausdorff distance, and prediction accuracy for both static fixation locations and dynamic scan-paths.

Keywords

---

*Corresponding author.

1. Introduction

Searching the localization of targets is still a challenge problem in the fields of computer vision. However, humans perform this task in a more intuitive and efficient manner by selecting only a few regions to focus on, while observers never form a complete and detailed representation of their surroundings [1]. Due to the high efficiency of this biological approach, more and more researchers are devoting increasingly great effort to probing the nature of attention [2].

Usually two kinds of top-down cues are used to predict human gaze location in dynamic scenes [3] and gaze movement control when searching target: cues about bottom-up features such as shape, color, shape, scale [4]-[7] and cues about the top-down visual context that contains the target as well as other relevant objects' spatial relationships and their environmental features [8]-[10].

In classical search tasks, target features are important source of guidance [11]-[15]. Although a natural object, such as an animal (cat or dog), does not have single defining feature, its statistically reliable properties (round head, straight body, legs and others) can be selected by visual attention. There has been little research using visual context in object search. Global context was used by Torralba to predict the region where the target is more detected by [16]. Object detectors are used by Ehinger and Paletta [17] [18] to search the targets in that predicted region detected by [16] for accurate localization. An extended object template containing local context is used by Kruppa and Santana to detect extended targets and infer the location of the targets via the ratio between the size of the target and the size of the extend template in [19]. Most of above methods are just only based on either global context or local context cues. However, Miao *et al.* proposed a serial of neural coding networks in [20]-[23] using both of them.

In this study, the main purpose of our work is to simulate human saccadic scan-paths by the proposed model in [23]. To test the performance of the proposed model, we collect human eye movement data by using an SMI iVIEW X Hi-Speed eye tracker on a face dataset with a sampling rate of 1250 Hz. We compare the saccadic scan-paths generated by the proposed model against actual human eye movement data from the face dataset [28].

The paper is organized as follows: the model of the gaze movement control in target searching proposed in [23] is introduced briefly in Section 2. In Section 3, we compare our saccadic scan-paths with previous methods and scan-paths from eye tracking data. Our conclusions are presented in Section 5.

2. Review of the Gaze Movement Control Model in Target Searching

This paper applies the target searching model in [23] to simulate the eye-motion traces. The feature used in the model is a kind of binary codes called Local Binary Pattern (LBP) [32], which has been proved through our work superior to orientation features used in the same system [33] [34] with respect to search performance. LBP is a simple and fast encoding scheme to map a 3×3 image patch to a local feature pattern in terms of an 8-bit code. This encoding scheme has no parameters to do such mapping, just outputting 0/1 for each bit through comparing the central pixel's value and that of each one of eight surrounding pixels. There are encoding and decoding parameters in the model [23], such as P, which determines how many context coding neurons are activated through competition. Through our experiments, we find the best value of 70% for this parameter. So in this paper, we use the best model with LBP feature and P = 70% to simulate eye-motion traces.
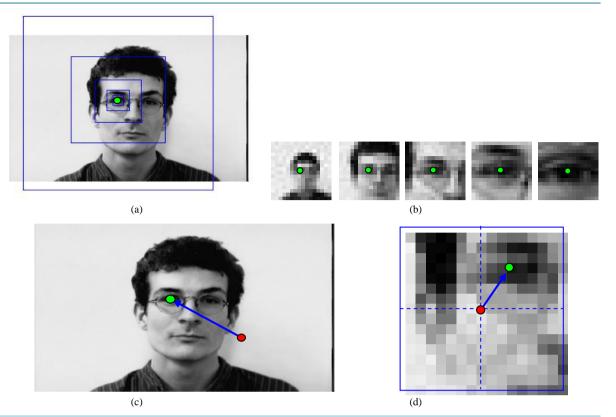
The learning and testing algorithm for target search is illustrated in **Figure 1** and described in Section 2.1 and 2.2. Here the visual context means the visual field image and the spatial relationship from the center of the visual field to the center of the target. In order to encode such context, we need to calculate and store the representation coefficients of the spatial relationship and the visual field images. The model's learning algorithm and test method are introduced in this part. In this experiment, we use head-shoulder image database from the University of Bern [24].

2.1. Model Training

The learning algorithm is described by [23] as follows:
1) Choose a value s from the scale set $\{s_j\}$ for the visual field that will be processed;
2) Choose an initial view point (x_j, y_j) as the center of the visual field from an initial point set $\{(x_j, y_j)\}$ covering the surrounding area of the target;

Figure 1. Illustration of learning and testing algorithm for target search. (a) Five visual fields centered at a gaze point (here is the left eye center); (b) Five visual field images (16×16 pixels, scales = 5, 4, 3, 2 and 1) sub-sampled from the original image (320×214 pixels) with intervals = 16, 8, 4, 2, 1 pixel(s); (c) The spatial relationship between one given starting gaze point and the target center; (d) Memorizing the visual context or predicting between the target center from current gaze points at different scales.

3) Receive signals from the current visual field, and output a relative position evaluation for the target with view point moving distances (Δx, Δy) ;

4) If the prediction error err is larger than the limit ERR(s) for the scale s of the current visual field, move the visual field center to a new position randomly; go to 3 until err $\leq$ ERR(s) or the iteration number is larger than a limit;

5) If err > ERR(s), generate a new VF-image encoding neuron (let its response $R_k = 1$); encode the visual context by calculating and memorizing the connecting weights $\{w_{ij, k}\}$ between the new VF-image encoding neuron and the feature neurons and the connection weights $w_{k,uv}$ between the new VF-image encoding neuron and the motion encoding neurons (let their response $R_{uv} = 1$) respectively using the Hebbian learning rule $\Delta w_{a,b} = \alpha R_a R_b$;

6) Go to 2 until all initial view points are chosen;

7) Go to 1 until all scales are chosen.

2.2. Model Prediction

In the test stage, the entire algorithm for view point control for object locating is given as follows:

1) Get a pre-given view point (x, y);

2) Choose a scale s from the set $\{s_i\}$ for the current visual field from the maximum to the minimum;

3) Receive signals from the current visual field, and calculating the response of the feature neurons and the context encoding neurons;

4) Predict a relative position (Δx, Δy) for the real position of the object;

5) If (Δx, Δy) = (0,0),object located;

6) If (Δx, Δy) $\neq$ (0,0), view point moving with (Δx, Δy), go 2 until all scales are chosen.

3. Experiments

3.1. Participants

Fifteen female and twelve male college students of Beijing University of Technology participated in this study. The age range was 23 - 26 and the average was 24 years old. All of the twenty-seven students had normal or corrected-to-normal vision.

3.2. Stimuli

A set of 30 face pictures are prepared as stimuli. Of this set of 30, 15 are Female-face, 15 are Male-face, and the size of each picture is 1024×768 pixels. Pictures are presented on a color computer monitor at a resolution of 1024 by 768 pixels. The monitor size was 41 cm by 33.8 cm, and the participants were sited in a chair about 76 cm in front of the screen.

Stimuli consist of a set of 30 face pictures. There are 15 Female-face and 15 Male-face in this set of 30, and each picture's size is 1024×768 pixels. One of the 30 face pictures are presented on a color computer monitor at a resolution of 1024 by 768 pixels.

3.3. Design

A new searching task was used in this study, participants were demanded to search the left and right eyes in a face from a pre-given starting point. Thirty pictures of face were used as stimuli, including 15 female and 15 male faces. The size of each picture was 1024×768 pixels. There were four pre-given starting points, named the first, second, third and fourth quadrant respectively in a counterclockwise direction, similar with those in a coordinated system. Searching from a starting point to a target eye decided the searching distance and direction. **Figure 1** illustrated the searching targets and the definition of the quadrants.

3.4. Procedure

For each trial, as shown in **Figure 2**, a black trail indicator was presented initially in the middle of the white screen for 1000 ms to indicate the target of the left or right eye. Then a "+" indicating pre-defined positions was presented in a random order. After that the picture of a face appeared in the middle of the screen for 2000 ms and participants were asked to search the right target eye or the left target eye as accurately and quickly as possible. Participants were told not to look at other part of the picture in the pictures after finding the target.
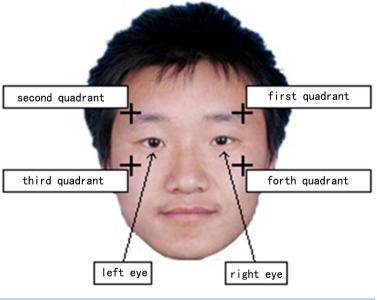

Figure 2. Sketch map of pre-given starting points in the face picture.

4. Experimental Results

4.1. Preprocess

The real fixation points are collected on the images with the size of 1024×768 pixels. However, the model can only deal with the gray images with the size of no more than 320×320 pixels. So when evaluating the performance of the model, we compress the original 1024×768 color images into 320×240 gray images. 10 face images are used in the learning stage and the other 20 face images are used for evaluation. The algorithms for the learning and the prediction stages are described in Section 2. When predicting fixation order and scan-paths, the same initial positions were used in the above experiment. Our model will search left eye and right eye separately from four different initial points that are similar to the above experiment. Each participant is asked to search left and right eyes from four different starting points on a face, and then it would certainly produce 8 eye scan-paths. For 27 subjects and 20 face images, $27 \times 20 \times 8$ scan-paths are totally recoded.
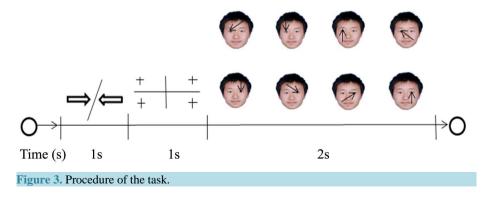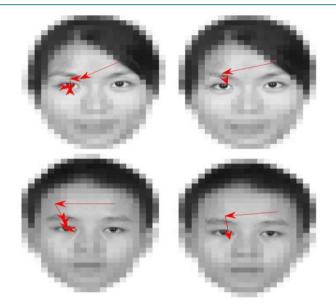
4.2. Evaluation of Fixation Order

We are aware of only a limited literature on computational models of active visual attention, and in particular active visual attention needs further investigation. Lee and Yu's work in [25] provided a conceptual framework but failed to provide a fully implemented solution with experimental results. Renninger *et al.* in [26] simulated scan-paths on novel shapes, but it is not clear how to adapt their method to natural images. However, Itti *et al.* in [27] proposed a scan path generation method from static saliency maps based on winner-takes-all (WTA) and inhibition-of-return (IoR) regulations. Tom Foul sham tried to find the evidence from normal and Gaze-Contingent search tasks in natural scenes in [28] for Itti. Marco Wischnewski proposed a model combining static and dynamic proto-objects in a TVA-based model of visual attention to predict where to look next in [29]. Gert Kootstra proposed a model to predict Eye Fixations on Complex Visual Stimuli Using Local Symmetry [30]. De Croon [31] proposed a novel gaze-control model, named act-detect, which use the information from local image samples in order to shift its gaze towards object locations for detecting objects in images. Our system can automatically generate the fixations, and the fixation can move to the target under the control of learned memory and experience in four or five steps. We here compare the simulated scan-paths generated by the model of [23] with human saccades. We select the initial positions on the four quadrants of the image shown in **Figure 3**. And the experimental results are illustrated in **Figure 4**. We can find that the simulated scan-paths by our model are similar to human saccades.

4.3. Distance of Scan-paths

In order to quantitatively compare the stochastic and dynamic scan-paths, we divide scan-paths into pieces of length 2. We use the Hausdorff distance to evaluate the scan-paths by the model proposed by Miao *et al.* with scan-paths of all subjects recorded by the eye tracker and evaluate the scan-paths between different subjects. The results are shown in **Table 1**.

In **Table 1**, Model-Human means the average of the Hausdorff distances between the scan path generated by model and that from each one of 27 subjects on corresponding images. Human-Human means the average of the Hausdorff distances between the scan-paths generated by any two of 27 subjects. We can know from **Table 1** that the simulated scan-paths by the model of Miao's are similar to human saccades by comparing the Hausdorff distance of scan-paths between the model and the humans: the average of the Hausdorff distances between

Figure 3. Procedure of the task.

Figure 4. The left column describes fixations predicted by the model proposed in [23]; the right column describes the real Fixations recorded by the SMI iVIEW X Hi-Speed eye tracker (Note: Here example face images are processed with mosaics).

Table 1. The average of the Hausdorff distances between the model to each one of 27 subjects and that between each pair of subjects.

Image	Hausdorff	
	Model-Human	Human-Human
1.bmp	25.83	23.38
2.bmp	26.88	26.06
3.bmp	20.23	23.49
4.bmp	29.02	27.77
5.bmp	28.37	26.84
6.bmp	27.70	25.90
7.bmp	27.94	31.44
8.bmp	28.06	23.50
9.bmp	33.95	30.39
10.bmp	32.11	30.76
11.bmp	25.40	21.67
12.bmp	28.68	27.41
13.bmp	32.90	28.01
14.bmp	30.42	30.18
15.bmp	29.73	22.97
16.bmp	30.60	28.50
17.bmp	30.44	15.46
18.bmp	31.44	26.21
19.bmp	34.39	36.96
20.bmp	29.46	20.29
Average	29.18	26.36

scan-paths generated by the model and each subject on all the corresponding images is 29.18 which is similar to the average (26.36) of the Hausdorff distances between the scan-paths generated by every two subjects of the total 27 subjects. We also compute the average of the Hausdorff distances in the cases of that the initial position is from the second, third and fourth quadrants respectively shown in **Table 2**.

In **Table 2**, Model-Human means the average of the Hausdorff distances between the scan-paths generated by the model and each of 27 subjects on all the corresponding images. Human-Human means the average of the Hausdorff distances between the scan-paths generated by every two of 27 subjects from the first, the second, the third and the fourth quadrants. The average of the Hausdorff distances from all four initial quadrants is 24.09. We conclude that the model of Miao's [23] achieves a good predictive accuracy on both static fixation locations and dynamic scan-paths.

4.4. Evaluation of Search Precision

We also compute the search precision from four different quadrants to left eye and right eye. The results are shown in **Table 3**. We noted that there is a discrepancy of the average value of the search precision between the left eye and right eye. Due to different contextual information which is coded and used by the search model, this case may take place.

5. Discussion and Conclusions

Miao *et al.* presented a new architecture for gaze movement control in target searching in [23]. This paper utilizes the model to simulate human saccadic scan-paths in target searching. To test the performance of the proposed model, we collect human eye movement data by using an SMI iVIEW X Hi-Speed eye tracker at a sample rate of 1250 Hz. We compare the saccadic scan-paths generated by the proposed model against human eye movement data. Experimental results demonstrate that the simulated scan-paths by the proposed model are similar to human saccades in terms of the fixation order and the Hausdorff distance of scan-paths. It can be learned that the model achieves good prediction accuracy on both static fixation locations and dynamic scan-paths.

The model is suitable for target searching in strong-context cases. However, it performs less effectively in weak-context cases. Thus as future work we hope to propose to use a bottom-up saliency map together with a top-down target template to assist context based object searching in weak context cases, in order to achieve good prediction accuracy on both static fixation locations and dynamic scan-paths in weak-context cases.

Table 2. The average of the Hausdorff distances.

Image	Hausdorff	
	Model-Human	Human-Human
First quadrant	29.18	26.36
Second quadrant	21.08	22.59
Third quadrant	18.96	22.10
Fourth quadrant	33.75	25.30
All average	25.74	24.09

Table 3. Search precision from four different quadrants to left eye and right eye.

Quadrant	Target	
	Left eye	Right eye
First quadrant	100%	95%
Second quadrant	100%	100%
Third quadrant	100%	100%
Fourth quadrant	85%	100%
Average	96.25%	98.75%

The current simulation is based on the model with the optimal features and parameters tuned from the real face data. How much do the variation of features and parameters affect the simulation is a valuable question to be investigated? Evaluating the model's performance on the pictures of people's face rather than real face is also an interesting question. These are what we will study in the future work.

Acknowledgements

References

[1] Rensink, R., O'Regan, K. and Clark, J. (1997) To See or Not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Sciences*, **8**, 368-373.

[2] Tsotsos, J.K., Itti, L. and Rees, G. (2005) A Brief and Selective History of Attention. In: Itti, Rees and Tsotsos, Eds., *Neurobiology of Attention*, Academic Press, Salt Lake City, xxiii-xxxii.

[3] Mital, P., Smith, T., Hill, R. and Henderson, J. (2011) Clustering of Gaze during Dynamic Scene Viewing Is Predicted by Motion. *Cognitive Computation*, **3**, 5-24. http://dx.doi.org/10.1007/s12559-010-9074-z

[4] Zelinsky, G., Zhang, W., Yu, B., Chen, X. and Samaras, D. (2006) The Role of Top-Down and Bottom-Up Processes in Guiding Eye Movements during Visual Search. *Advances in Neural Information Processing Systems*, Vancouver, 5-8 December 2005, 1407-1414.

[5] Milanese, R., Wechsler, H., Gil, S., Bost, J. and Pun, T. (1997) Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Hilton Head, 21-23 June 1994, 781-785.

[6] Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N. and Nuflo, F. (1995) Modeling Visual Attention via Selective Tuning. *Artificial Intelligence*, **78**, 507-545. http://dx.doi.org/10.1016/0004-3702(95)00025-9

[7] Navalpakkam, V., Rebesco, J. and Itti, L. (2005) Modeling the Influence of Task on Attention. *Vision Research*, **45**, 205-221. http://dx.doi.org/10.1016/j.visres.2004.07.042

[8] Chun, M. and Jiang, Y. (1998) Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology*, **36**, 28-71. http://dx.doi.org/10.1006/cogp.1998.0681

[9] Chun, M. (2000) Contextual Cueing of Visual Attention. *Trends in Cognitive Sciences*, **4**, 170-178. http://dx.doi.org/10.1016/S1364-6613(00)01476-5

[10] Henderson, J., Weeks Jr., P. and Hollingworth, A. (1999) The Effects of Semantic Consistency on Eye Movements during Complex Scene Viewing. *Journal of Experimental Psychology*: *Human Perception and Performance*, **25**, 210-222. http://dx.doi.org/10.1037/0096-1523.25.1.210

[11] Treisman, A. and Gelade, G. (1980) A Feature Integration Theory of Attention. *Cognitive Psychology*, **12**, 97-136. http://dx.doi.org/10.1016/0010-0285(80)90005-5

[12] Wolfe, J.M. (1994) Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulletin & Review*, **1**, 202-238. http://dx.doi.org/10.3758/BF03200774

[13] Wolfe, J.M. (2007) Guided Search 4.0: Current Progress with a Model of Visual Search. In: Gray, W., Ed., *Integrated Models of Cognitive Systems*, Oxford Press, New York. http://dx.doi.org/10.1093/acprof:oso/9780195189193.003.0008

[14] Wolfe, J.M., Cave, K.R. and Franzel, S.L. (1989) Guided Search: An Alternative to the Feature Integration Model for Visual Search. *Journal of Experimental Psychology*: *Human Perception and Performance*, **15**, 419-433. http://dx.doi.org/10.1037/0096-1523.15.3.419

[15] Zelinsky, G.J. (2008) A Theory of Eye Movements during Target Acquisition. *Psychological Review*, **115**, 787-835. http://dx.doi.org/10.1037/a0013118

[16] Torralba, A. (2003) Contextual Priming for Object Detection. *International Journal of Computer Vision*, **53**, 169-191. http://dx.doi.org/10.1023/A:1023052124951

[17] Ehinger, K., Hidalgo-Sotelo, B., Torralba, A. and Oliva, A. (2009) Modelling Search for People in 900 Scenes: A Combined Source Model of Eye Guidance. *Visual Cognition*, **17**, 945-978. http://dx.doi.org/10.1080/13506280902834720

[18] Paletta, L. and Greindl, C. (2003) Context Based Object Detection from Vide*o*. In: *Proceedings of International Conference on Computer Vision Systems*, Graz, 502-512. http://dx.doi.org/10.1007/3-540-36592-3_48

[19] Kruppa, H., Santana, M. and Schiele, B. (2003) Fast and Robust Face Finding via Local Context. In: *Proceedings of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Nice, France, 11-12 October, 2003, 1-8.

[20] Miao, J., Chen, X., Gao, W. and Chen, Y. (2006) A Visual Perceiving and Eyeball-Motion Controlling Neural Network for Object Searching and Locating. *Proceedings of International Joint Conference on Neural Networks*, Vancouver, 4395-4400.

[21] Miao, J., Zou, B., Qing, L., Duan, L. and Fu, Y. (2010) Learning Internal Representation of Visual Context in a Neural Coding Network. *Proceedings of the International Conference on Artificial Neural Networks*, Thessaloniki, 15-18 September 2010, 174-183. http://dx.doi.org/10.1007/978-3-642-15819-3_22

[22] Miao, J., Qing, L., Zou, B., Duan, L. and Gao, W. (2010) Top-Down Gaze Movement Control in Target Search Using Population Cell Coding of Visual Context. *IEEE Transactions on Autonomous Mental Development*, **2**, 196-215.

[23] Miao, J., Duan, L., Qing, L. and Qiao, Y. (2011) An Improved Neural Architecture for Gaze Movement Control in Target Searching. *Proceedings of the IEEE International Joint Conference on Neural Networks*, San Jose, 31 July-5 August 2011, 2341-2348.

[24] The Face Database of the University of Bern (2008). http://www.iam.unibe.ch/fki/databases/iam-faces-database

[25] Lee, T. and Yu, S. (2002) An Information-Theoretic Framework for Understanding Saccadic Eye Movements. *Advanced in Neural Information Processing System*, **12**, 834-840.

[26] Renninger, L., Verghese, P. and Coughlan, J. (2007) Where to Look Next? Eye Movements Reduce Local Uncertainty. *Journal of Vision*, **7**, 6.

[27] Itti, L., Koch, C. and Niebur, E. (1998) A Model of Saliency Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 1254-1259.

[28] Foulsham, T. and Underwood, G. (2011) If Visual Saliency Predicts Search, Then Why? Evidence from Normal and Gaze-Contingent Search Tasks in Natural Scenes. *Cognitive Computation*, **3**, 48-63. http://dx.doi.org/10.1007/s12559-010-9069-9

[29] Wischnewski, M., Belardinelli, A., Schneider, W. and Steil, J. (2010) Where to Look Next? Combining Static and Dynamic Proto-Objects in a TVA-Based Model of Visual Attention. *Cognitive Computation*, **2**, 326-343. http://dx.doi.org/10.1007/s12559-010-9080-1

[30] Kootstra, G., de Boer, B. and Schomaker, L. (2011) Predicting Eye Fixations on Complex Visual Stimuli Using Local Symmetry. *Cognitive Computation*, **3**, 223-240. http://dx.doi.org/10.1007/s12559-010-9089-5

[31] de Croon, G., Postma, E. and van den Herik, H. (2011) Adaptive Gaze Control for Object Detection. *Cognitive Computation*, **3**, 264-278. http://dx.doi.org/10.1007/s12559-010-9093-9

[32] Ojala, T., Pietikainen, M. and Harwood, D. (1996) A Comparative Study of Texture Measures with Classification Based on Featured Distribution. *Pattern Recognition*, **29**, 51-59. http://dx.doi.org/10.1016/0031-3203(95)00067-4

[33] Miao, J., Chen, X., Gao, W. and Chen, Y. (2006) A Visual Perceiving and Eyeball-Motion Controlling Neural Network for Object Searching and Locating. *Proceedings of the International Joint Conference on Neural Networks*, Vancouver, 16-21 July 2006, 4395-4400.

[34] Miao, J., Duan, L.J., Qing, L.Y., Gao, W. and Chen, Y.Q. (2007) Learning and Memory on Spatial Relationship by a Neural Network with Sparse Features. *Proceedings of the International Joint Conference on Neural Networks*, Orlando, 12-17 August 2007, 1-6. http://dx.doi.org/10.1109/ijcnn.2007.4371293