Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

# Office-mate: Selective attention and incremental object perception

Minho Lee, Young-Min Jnag
Department of Electrical Engineering and Computer Science
Kyungpook National University
1370 Sankyuk-Dong, Puk-Gu, Taegu 702-701, Korea
mholee@knu.ac.kr, ymjang@ee.knu.ac.kr

Sang-Woo Ban
Department of Information and Communication Engineering
Dongguk University
707 Seokjang-Dong, Gyeongju, Gyeongbuk 780-714, Korea
swban@dongguk.ac.kr

***Abstract*— We propose an autonomous robot vision system that is applied to develop an intelligent artificial officemate. In order to operate the proposed system in real environment, it is very important for the officemate to be able to adapt to an environmental changes that may occur in an indoor environment. Novelty detection is one of essential functions for the officemate to detect a situation change. The proposed system can indicate a novel scene and a scene change based on a visual selective attention module. Moreover, it can adaptively acquire new information based on incremental object perception, face recognition, and emotion representation. In order to implement an on-line officemate system, we implement an efficient model by simplification and optimization procedure which can reduce the computation load. Experimental results show that the developed system successfully identifies a change of natural scenes and incrementally learns an arbitral object and a face, and it can also extend its knowledge through interaction with human supervisor.**

***Keywords*—stereo saliency map, object perception, face recognition, emotion representation, office-mate, autonomous robot**

## I. INTRODUCTION

The present intelligent system is mainly working based on task-specific manners and/or the programmer's input about environmental information. Even though such a traditional artificial intelligent robot has been successfully used for simple tasks, a new artificial intelligent paradigm is needed to continuously adapt the changes of an environment and to voluntarily increase new knowledge without any interruption by programmers [1 - 3]. Recently, a new concept, so-called Autonomous Mental Development (AMD) [1 - 3] has been proposed for the construction of a more intelligent robot. The AMD mechanism is that robots can increase their own knowledge through their interaction with the environment, just as humans do. In order to develop the intelligent robot with the AMD concept, we need to implement more human-like sensors such as retina, electronic nose, touch, smell and acoustic sensors. Additionally, we need to develop an intelligent model in order to pay attention to interesting objects by primitive sensory information. Furthermore, it is important that humans and the environment can interactively share their knowledge.

In this paper, we consider to develop a vision based officemate system [4] that can autonomously understand office environment and human's emotion, through which it can play a role for a mate of an office worker by helping a visitor in the office. In order to implement such an officemate robot system, we need to consider many complex functions. However, in this paper, we consider only a simple officemate system with limited functions that can identify a novel situation whether the environment has changed compared to the previous states. Also, we consider the environmental recognition functions that perceive an arbitrary object existing in visual field around the developed system [5 - 7]. The proposed system is based on the human-like selective attention model that imitates the human visual search mechanism in a context-free environment in order to focus on an interesting object. The proposed selective attention model is implemented by considering the visual pathway in our brain [8].

Moreover, one of the most important requisites for the officemate system is to have a robust object perception model with the capability to generally represent an arbitrary object as the human vision system does. The performance of the object representation model highly depends on the robust feature extraction method. Therefore, we propose an incremental hierarchical MAX (IHMAX) [9] model for incremental object perception and memorization, which can construct both object color and form feature clusters.

In addition, face recognition and emotion representation are also important features for an officemate system. In this paper, we consider an eigen feature based face recognition model using an incremental principal component analysis (IPCA) in conjunction with an one-class support vector machine (OCSVM). Moreover, an active appearance model (AAM) and OCSVM are applied to indicate emotion status based on facial expressions. The proposed autonomous robot vision system was implemented in order to work in real environment.

We describe the proposed officemate system consisting of a selective attention module, a novelty detection module,



978-1-4244-2794-9/09/$25.00 ©2009 IEEE

an object-perception module, a face-recognition module, and an emotion representation module in Section Ⅱ. In section Ⅲ, we describe the implementation of proposed autonomous robot vision system and the experiment results. Conclusion and discussion will follow in the final section.

## II. THE PROPOSED OFFICE-MATE SYSTEM

Fig. 1 shows an overview of the proposed officemate system. The proposed officemate system consists of five modules such as a visual selective attention module, a novelty scene detection module, an object perception module, a face recognition module, and an emotion representation module.

The visual selective attention model can be efficiently used for sequential visual search process as a preprocessor. The visual selective attention model has bottom-up mechanisms for deciding salient areas. As a bottom-up manner, the visual selective attention model can decide a salient area by calculating the relativity of primitive visual features such as intensity, edge, and color.

The proposed officemate system can incrementally extend and enhance its knowledge about scenes under an office environment by indicating a novel scene and acquiring new object information from a novel scene. The novelty scene detection model works for indicating novelty or changes in a scene obtained in an office environment, which uses a saliency map generated by the attention model as a scene representation [10, 11]. The novelty scene detection model also works for the officemate system to efficiently process complex visual information in an office environment by treating novel scenes only.

Moreover, in order to develop an incremental object non-specific perception model, we considered an IHMAX model. Moreover, the proposed system can catch and represent human emotion based on understanding human facial expression as well as human face recognition.

The developed officemate system can also interface with a text-to-speech module to make an appropriate responsive action for human beings or environment.

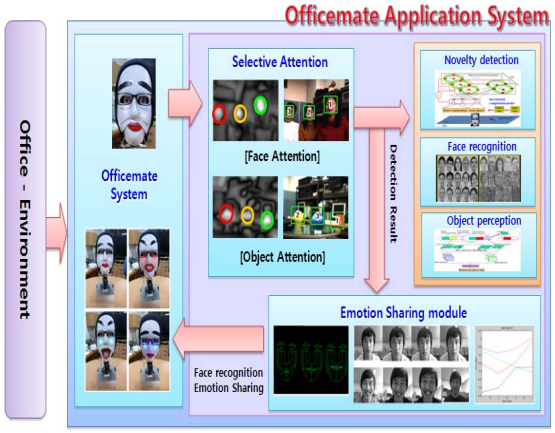


Figure 1. Overview of the proposed officemate system.

### A. *Stereo Saliency Map Model for Selective Attention*

Fig. 2 illustrates the proposed stereo saliency map model for selective attention, which is partly inspired by biological visual pathway from the retina to the visual cortex through the lateral geniculate nucleus (LGN) for bottom-up processing [12, 13]. In order to implement a visual attention function, two processes are combined to generate a saliency map. One is to generate two bottom-up saliency maps from the left image and the right image, respectively. The other is to generate the final stereo saliency map by integrating the bottom-up saliency map and depth information.

In this model, selective attention regions in each camera are obtained from each saliency map, and are then used for selecting a dominant landmark.

The saliency map is generated by calculating the relativity of primitive visual features such as intensity, edge, and color as shown in Fig. 2. And, based on the single eye alignment hypothesis [14], comparing the maximum salient values within selective attention regions in two camera images, we can adaptively decide the master eye that has a camera with a larger salient value [12, 13]. After successfully localizing corresponding landmarks on both left and right images with master and slave eye, we are able to get depth information by simple triangulation [12, 13]. The stereo saliency map uses the depth information specifically, in which the distance between a camera and a focused region is used as a characteristic feature in deciding saliency weights.

### B. *Novelty Scene Detection*

In order to develop a low level novelty scene detection model, we developed a robust scene description model with a tolerance against noise and a slight affine transformation such as translation, rotation, zoom-in and zoom-out in a dynamic environment. Such a model can be used to detect novelty by comparing the description of the current scene with that of an experienced scene, like the assumable function of the hippocampus in our brain.

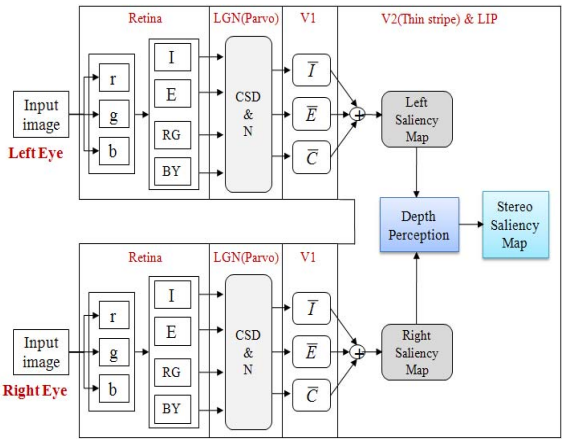


Figure 2. The stereo saliency map model (r: red, g: green, b: blue, I: intensity image, E:edge image, RG: red-green opponent coding image, BY: blue-yellow opponent coding image, CSD & N: center-surround difference and normalization, $\overline{I}$ : intensity FM, $\overline{E}$ : orientation FM, $\overline{C}$ : color FM)

Fig. 3 shows the proposed model, each scene is represented using salient area information obtained from the bottom-up SM model such as the geometric topology of the salient areas, degree of saliency and size of each salient area. Therefore each scene is represented by three components such as a relative distance score $f_d$, a relative scale score $f_s$, and an energy score $f_e$. A relative distance score $f_d$ is calculated using relative distance of the salient areas from the decided reference salient area. A relative scale score $f_s$ is obtained from relative scales of the salient areas. And, an energy score $f_e$ is calculated using degree of saliency of the salient areas. The algorithm for these three components is described in [10].

As shown in Fig. 2, the obtained three components for representing an input scene are used as the input of the developed incremental computation model, which can incrementally memorize scene information. Moreover, the proposed model is able to recognize the placement and the direction of a field of view from memorized scene information in the incremental computation model.

If the model is executed for the first time, the incremental computation model needs to be initiated with only one output neuron. When the input of the incremental computation model is prepared, the proposed model obtains the location information $p_i$ of the camera system and the direction information $d_j$ of the field of view of the camera system, by feedback information such as encoder sensors. If the location information is new, the incremental computation model generates a new output node for that location and direction. After getting the location and direction information, the proposed model selects a winner node from among the nodes related with the dedicated location and direction [10].

Then, the model checks the degree of similarity using a vigilance value $\rho'$. If the degree of similarity between the current input scene and the winner node is less than the vigilance value $\rho'$, then the proposed model indicates it is a novelty scene and creates a new node, otherwise the current input scene is considered as previously seen and the proposed model adjusts the weights of the winner node according to the input scene through the unsupervised learning [10].

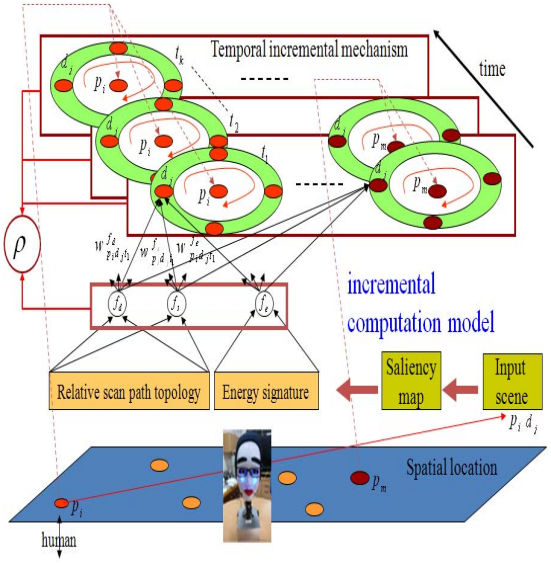


Figure 3. Incremental model for recognizing an environment

The developed model can recognize the location of its placement by using the trained incremental computation model for scene memorization and scene novelty detection, as described above. The proposed placement recognition algorithm uses the same procedure that makes an input vector $(f_d, f_s, f_e)$ for the incremental computation model [10].

### C. *Incremental Object Perception Model*

Most of vision system models are separated by training process and test process. During training process, the vision system learns sample data with limitation of circumstance in a database. That kind of system works well under similar conditioned circumstance. In unfamiliar places, however, it may not perform well without incremental adaptation characteristic. The reason is that the parameters (or prototypes) in the model are obtained from specific conditioned environment.

The conventional HMAX model uses static prototypes which are made during off-line training with natural scenes. These prototypes have a limitation for adaptation to unknown environment. The static 2,000 prototypes for object representation were made by randomly selected location and random scale band from C1 feature map in the HMAX model using natural scenes. In the sequel, these static prototypes have restricted or poor sparseness property. Therefore, we propose a new model that has an incremental characteristic for being able to adapt to unknown environment and generates more plausible prototypes [15, 16].

Fig. 4 shows the proposed incremental object perception model including feature extraction and an incremental object representation part [9].

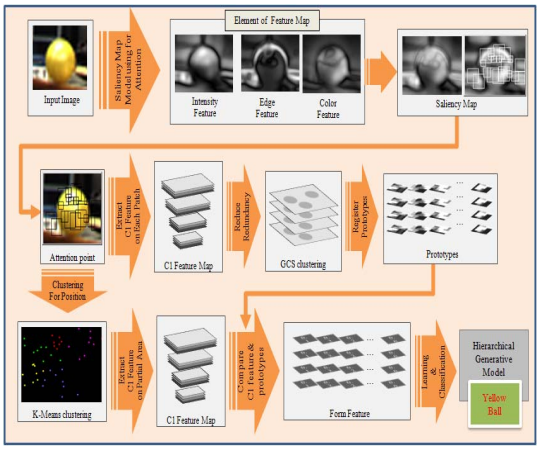


Figure 4. Outline of proposed incremental object perception model

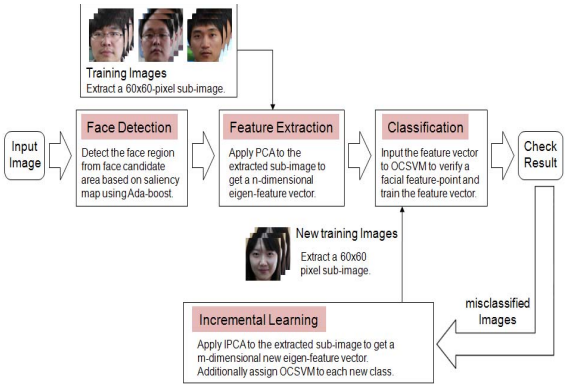


Figure 5. Incremental face recognition model

In the proposed model, we use the C1 feature and dynamic prototypes obtained by a growing cell structure (GCS) and the bottom-up SM model. The C1 feature is known as having robust characteristic about scale and translation. The GCS [17] can give incremental prototypes as making clusters about similar nodes based on topology information of input data.

### *D. Incremental Face Recognition Model*

The proposed face recognition system mainly consists of the following functional modules: face detection, face recognition and incremental learning. The operations in these modules must be done online without any human intervention. Fig. 5 shows the overall process in the proposed system. In the face detection part, at each time frame, face candidate areas are localized based on saliency map using Adaboost.

In each operation, a small sub-image is extracted from a candidate area, then the eigenfeatures of the sub-image are given to a OCSVM to verify if it corresponds to any one of the facial features. Those eigen features are obtained using the PCA. Next, in the incremental face recognition part, some rectangular regions of the face candidate areas can be extracted from the new image, and then each of the extracted regions is transformed into an eigen feature vector using IPCA [18] in order to reduce the dimensions. This eigenfeature is given to OCSVM for feature training, and then a recognition result is obtained. The misclassified images are collected to use as training samples for incremental learning.

### *E. Emotion Representation using Facial Expression Model*

Facial expressions provide information about human affective status. In order to analyze the emotional information from the human facial expressions, the AAM [19] is used to extract relevant information regarding the shapes of the faces to be analyzed. Specific key points from a facial characteristic point (FCP) model are employed to derive the set of features, which are used as one of the emotional features for an OCSVM.

The shape information is used to compute a set of parameters that describe appearance of the facial features.

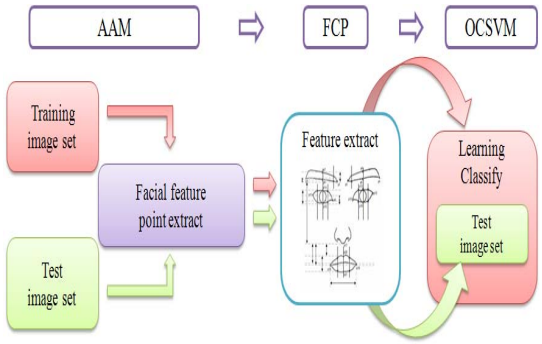


Figure 6. Emotion recognition model based on understanding facial expression.

The AAM models the shape and texture information. To identify specific features for the facial expression recognizer, we need to select the optimal key points on the face area from the shape data. The key points are defined as facial characteristic points (FCPs) [20]. Fig. 6 shows emotion recognition based on facial expression.

## III. HARDWARE IMPLEMENTATION & EXPERIMENTS

We implemented a stereo vision robot unit for an autonomous robot vision system for recognizing a visual environment. Fig. 7 shows the implemented system called by SMART-v2.0 (Self Motivated Artificial Robot with a Trainable selective attention model version 2.0). The SMART-v2.0 has seven DOF and two USB web cameras and a Text to Speech module (TTS) to communicate with humans to inquire about an interesting object, and tilt sensor to set offset position before starting moving. We use the Atmega128 as the motor controller and zigbee to transmit motor command from a PC to SMART-v2.0. The SMART-v2.0 can search an interesting region by the selective attention model, which is explained in section Ⅱ. Fig. 8 shows the platform of the SMART-v2.0 and its working system configuration.

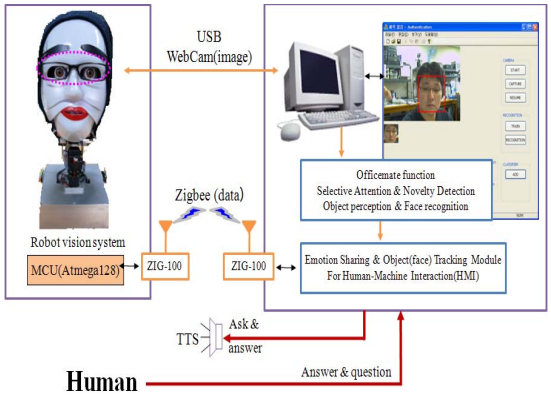


Figure 7. SMART-v2.0 platform

Fig. 8 shows the emotion expression of SMART-v2.0, at neutral, sad, happy, angry, and surprise. Additionally, it can detect a novel scene in an active camera system, not in a static camera system. Fig. 9 shows that the proposed model of officemate system demonstration flows. Fig. 10 shows the experimental results of emotion recognition of the SMART-v2.0 system. As shown in Fig. 10, the proposed model successfully recognizes 4 different kinds of emotions.

Following is an actual working scenario of the proposed officemate system based on human-machine interaction:

1) The SMART-v2.0 is set in a mobile phone shop. The SMART-v2.0 has a priori knowledge about the environment and can indicate environmental changes.

2) The SMART-v2.0 can indicate a customer who enters into the office by itself. The customer is indicated by the novelty scene detection model in conjunction with the face recognition model. For example, the SMART-v2.0 can say 'hello' to the customer and the customer can ask environment information such as a location of a specific object to the SMART-v2.0.

3) The proposed model can track face locations using the face recognition model. Moreover, a vergence control of the camera system is working to focus on a face with depth information obtained from the stereo SM model.

4) The SMART-v2.0 memorizes a customer's face information localized at the selected regions by the incremental face recognition model.

5) Face pose estimation function is also working in the proposed system in order to detect the customer's intention based on indicating the direction of turned customer's face.

6) Then, an incremental object perception using the IHMAX and hierarchical generative model is conducted by the proposed system in order to indicate the object that the customer pays attention to.

7) The officemate system is trying to understand and represent emotion of the customer for catching the customer's preference for the attended object such as mobile phone goods selling in the shop and it can follow the customer's facial expression. Moreover, the system provides information for the goods to customer.

8) Through which, the officemate system can help the shop worker based on human-like visual perception and interaction.


Figure 8. Emotion expression of SMART-v2.0

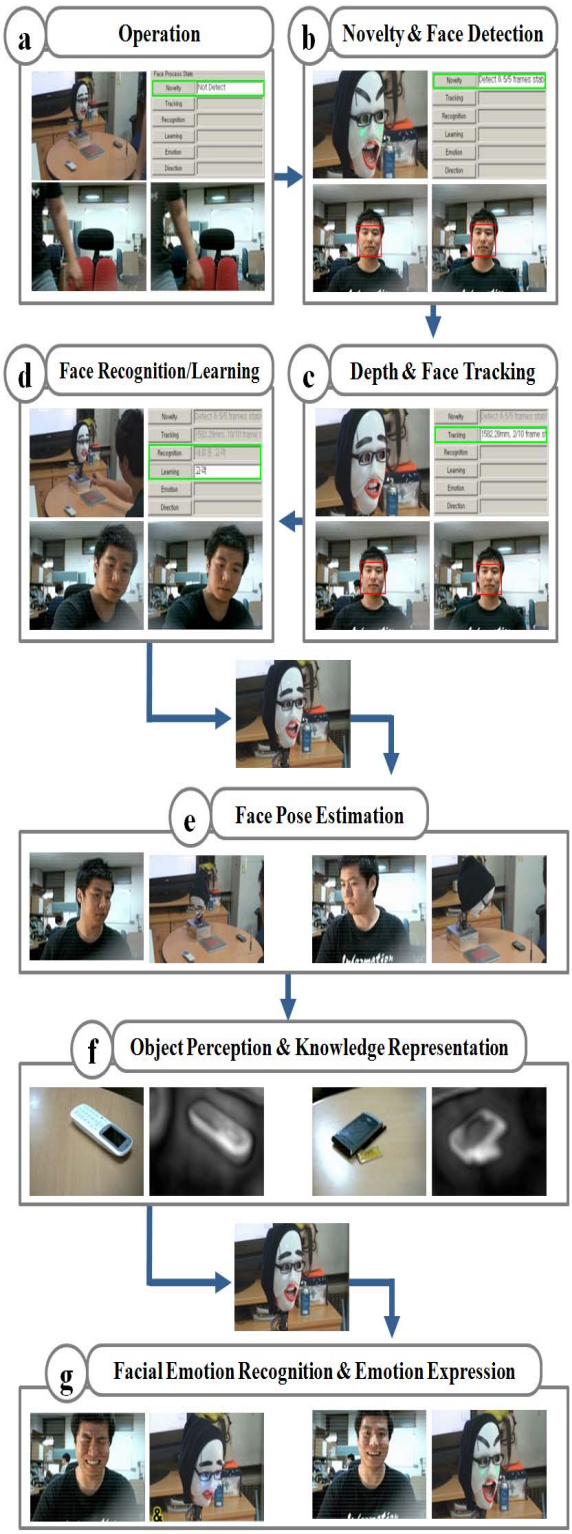


Figure 9. Demonstration flows the officemate system

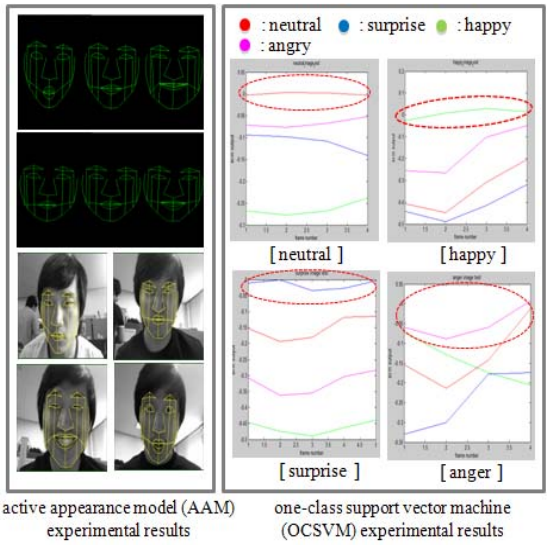


Figure 10. The experimental results of the emotion recognition

## IV. CONCLUSION

We developed and implemented a novel autonomous robot vision system that can operate as an officemate system, which can incrementally understand visual environment by using the five functions: visual selective attention, object perception, face recognition, emotion representation, and novelty scene detection. The incremental model is based on novelty detection, and successfully recognizes both new environment and change of environment. The proposed model can represent and discriminate more and more objects incrementally by the generative object perception model using the IHMAX, and it can recognize a class of an object based on hierarchical generative model. The essential roles such as face recognition and emotion sharing with visitors and office workers are successfully done by the incremental face recognition model and emotion representation.

The developed system can play a role as an office mate for office workers and office visitors by providing them with helps through human interaction. Our future research will concentrate on developing a more advanced officemate system with knowledge representation and reasoning.

## REFERENCES

[1] M. Asada, K. F. MacDorman, H. Ishiguro, Y. Kuniyoshi, "Cognitive developmental as a new paradigm for the design of humanoid robots," Robotics and Autonomous Systems, vol. 37, pp. 185 – 193, 2001.

[2] J. Weng, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen, "Autonomous mental development by robots and animals," Science, vol. 291, pp. 599-600, 2001.

[3] C. Breazeal, "Designing Social Robots," MIT Press, Cambridge, MA, 2002.

[4] W. J. Won, S.W. Ban, and M. Lee, "Self-motivated autonomous robot with a trainable slelctive attention model," Intelligent Automation and Soft Computing, vol. 15, no 2, pp. 1-16, 2009

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

[6] S. J. Park, K. H. An and M. Lee, "Saliency map model with adaptive masking based on independent component analysis," Neurocomputing, vol. 49, pp. 417-422, 2002.

[7] W. J. Won, J. Y. Yeo, S. W. Ban, and M. Lee, "Biologically motivated incremental object perception based on selective attention," IJPRAI, vol. 49, pp. 417-422, 2002.

[8] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Patt. Mach. Intell, vol. 20, no. 11, pp. 1254-1259, 1998.

[9] M. H, Kim, S. W. Ban, and M. Lee, "Biologically Motivated Object Perception Using Incremental Hierarchical MAX model," ICONIP2008.

[10] S. W. Ban and M. Lee, "Autonomous incremental visual environment peception based on visual selectie attention," IJCNN, pp. 1411-1416, 2007.

[11] S. W. Ban and M. Lee, "Selective attention-based novelty scene detection in dynamic environments," Neurocomputing, vol. 69, no.13-15, pp. 1723-1727, 2006.

[12] S. Jeong, S. W. Ban, and M. Lee, "Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment", Neural Networks, vol. 21, pp.1420-1430, 2009.

[13] S. B. Choi, B. S. Jung, S. W. Ban, H. Niitsuma, M Lee, "Biologically motivated vergence control system using human-like selective attention model," Neurocomputing, vol. 69, pp. 537-558, 2006

[14] F. Thorn, J. Gwiazda, A. A. Cruz, J. A. Bauer, and R. Held, "The development of eye alignment, convergence, and sensory binocularity in young infants," Investigative Ophthalmology and Visual Science, vol. 35, pp. 544_553, 1994.

[15] M. Riesenhuber, and T. Poggio, "Hierachical models of object recognition in cortex," Nature Neuroscience, vol. 2, pp. 1019–1025, 1999.

[16] C. Cadieu, M. Kouh, A. Pasupathy, C. E. Connor, M. Riesenhuber and T. Poggio, "A model of V4 shape selectivity and invariance," Neurophysiol, vol. 98, pp.1733-1750, 2007.

[17] S. Marsland, J. Shapiro, and U. Nehmzow, "A self-organising network that grows when required," Neural Networks, Special Issue vol. 15, pp. 1041–1058, 2002.

[18] S. Ozawa, S. L. Toh, S. Abe, S. Pang and N. Kasabov, "Incremental learning for online face redognition," Proceedings of International Joint Conference on Neural Networks, vol. 18, pp. 575-584, 2005.

[19] A. U. Batur and M. H. Hayes, "Adaptive active appearance models," IEEE Trans. Image Processing, vol. 14, pp. 1707–1721, 2005.

[20] H. Kobayashi, F. Hara, "Facial interaction between snimated 3D face robot and human beings," IEEE Computer Society Press, pp. 3732–3737, 1997