한국마린엔지니어링학회지 제36권 제5호, pp. 674~682, 2012. 7(ISSN 2234-8352 Online) / (ISSN 2234-7925 Print)
Journal of the Korean Society of Marine Engineering http://dx.doi.org/10.5916/jkosme.2012.36.5.674

Implementation of a Stereo Vision Using Saliency Map Method

Hyeung-sik Choi[1] · Hwan-sung Kim† · Hee-Young Shin[2] · Min-ho Lee[3]

(Received April 26, 2012; Revised June 12, 2012; Accepted July 6, 2012)

Abstract: A new intelligent stereo vision sensor system was studied for the motion and depth control of unmanned vehicles. A new bottom-up saliency map model for the human-like active stereo vision system based on biological visual process was developed to select a target object. If the left and right cameras successfully find the same target object, the implemented active vision system with two cameras focuses on a landmark and can detect the depth and the direction information. By using this information, the unmanned vehicle can approach to the target autonomously. A number of tests for the proposed bottom-up saliency map were performed, and their results were presented.

Key words: Sensor system, Active stereo vision system, Saliency map, Unmanned vehicle

1. Introduction

A vision system is an important sensor for the unmanned vehicle such as the unmanned forklift for approaching destination. In order for the unmanned forklift to move with loads without an human beings, they need a number of sensors such as the gyro sensor, supersonic wave sensor, infrared rays sensor, and vision sensor. Out of the sensors, one of the most important sensor is the vision sensor. A number of researches were performed in application of vision sensor to the unmanned forklift [1,2]. However, the vision sensors in the research are single vision system such that it should be used with other sensors to find depth information.

Many researchers have studied behaviors of the human vision system to develop the intelligent vision system for many systems, but those schemes were not applied to the unmanned vehicle yet. When human eyes look a natural scene, left and right eyes converge on an interesting object by the action of the brain function and eyeball. This vergence mechanism is very effective in processing high dimensional data with great complexity. Bernardino and Victor implemented VCSS (vergence control stereo system) using log-polar images based on a simple retinal operation [3]. Batista et al. made VCSS using a retinal optical flow disparity and the target depth velocity [4]. But this system converges to the moving object because of using optical flow based on the motion information of the retina. These two approaches are insufficient to consider the brain function related with the vergence control.

Conradt et al. proposed a stereo vision system using a biologically inspired saliency map (SM) [5]. They detect landmarks in both images with interactions between the feature detectors and a simple SM, and finally obtain their direction and distance. Their proposed model does not fully consider the operation of the retina because they only consider the roles of neurons in the

† Corresponding Author(Korea Maritime university, Tel: 051-410-4334, E-mail: kimhs@hhu.ac.krr)
1 Korea Maritime University, Tel: 051-410-4297, E-mail: hchoi@hhu.ac.kr
2 Korea Maritime university, Tel: 051-410-4969, E-mail: vakarin@naver.com
3 Kyungpook National university, Tel: 053-950-6436 E-mail: mholeeknu.ac.kr

hippocampus mainly related with depth information.

We focus on a new active vision system with the human-like vergence function. In section 2, a sensor fusion system is explained. Especially, the algorithm and theory of the active stereo vision sensor is explained in detail. The control system for the unmanned vehicle including the sensor fusion system are explained in section 3. Finally, tests of the developed vision system for the unmanned vehicle are performed and results are discussed.

2. An analysis on the stereo vision sensor system

2.1 Structure of a stereo vision system

A stereo vision sensor system can be applied to motion and distance control and obstacle avoidance of intelligent autonomus vehicles such as an unmanned forklift. For these, stereo vision sensor system using two CCD camera module is developed. The stereo vision sensor system is composed of two cameras capable of yawing and pitching motion as shown in **Figure 1**. Each camera has two degree-of-freedom (DOF) and the base joint has one DOF such that the vision system has three DOF and can perform visual tracking of the object. The motors for the stereo CCD cameras are 3W and their controller is one chip microprocessor. The encoder on the motor senses the rotating position and speed of the motor. The stereo camera is used to find the depth information between the unmanned vehicle and an object. In this article, a human-like searching method for the depth information was proposed. The detail analysis for the active vision system is presented in the following section.

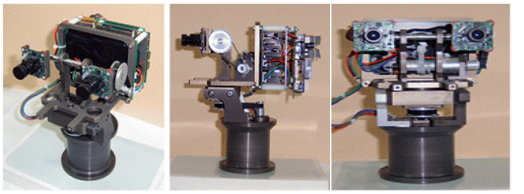


Figure 1: A picture of the stereo vision system

2.2 Bottom-up saliency map

Figure 2 shows a simple biological visual pathway from the retina to the visual cortex through the lateral geniculate nucleus (LGN). The photo receptor in the retina transforms an optical signal into an electrical signal. The transformed signals for the static image such as edge, intensity and color opponent information are transmitted to the visual cortex through the LGN. A Sobel operator was used to implement the edge extraction of the retina cell. In order to implement the color opponent coding, four broadly-tuned color channels were created: R = r - (g+ b)/2 for red, G = g - (r + b)/2 for green, B = b - (r + g)/2 for blue, and Y = (r + g)/2 - |r - g|/2 - b for yellow where r, b and g denote red, blue and green pixel values, respectively. RG and BY color opponent coding was obtained by considering the on-center and off-surround mechanism. Additionally, we used the noise tolerant generalized symmetry transform (NTGST) algorithm to extract symmetrical information from the edge information [6].

Figure 2 shows a proposed saliency map model based on biological visual process. The extracted visual information such as edge, intensity, and color opponency is transmitted to the visual cortex through the LGN in which symmetrical information can be extracted by edge information. That extracted information.

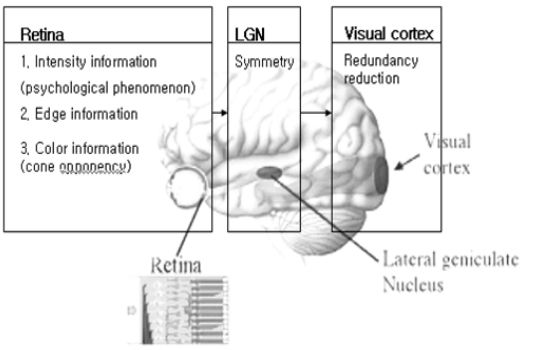


Figure 2: Biological visual pathway

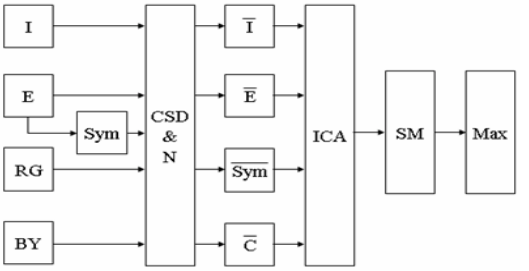


Figure 3: Bottom-up saliency map model

I: intensity feature, E: edge feature, Sym: symmetry feature, RG : red-green opponent coding feature map, BY: blue-yellow opponent coding feature, CSD & N : center-surround difference and normalization, $\overline{I}$: intensity feature map, $\overline{E}$: edge feature map, $\overline{Sym}$: symmetry feature map, $\overline{C}$: color feature map, ICA : independent component analysis, SM : saliency map , Max : max operator

The bottom-up saliency map model(**Figure 3**) is preprocessed to mimic the biological mechanism. In the course of preprocessing, we used a Gaussian pyramid with different scales from 0 to n level, in which each level is made by subsampling of 2n, thus constructing five feature bases such as I(∟), E(∟), Sym(∟), RG(∟), and BY(∟). It is to reflect the non-uniform distribution of retinotopic structure. Then, the center-surround mechanism is implemented in the model as the difference between the fine and coarse scales of Gaussian pyramid images [6, 9]. Consequently, five center-surround feature basis such as I(c,s), E(c,s), Sym(c,s), RG(c,s), and BY(c,s) are obtained by the following equations.

$$I(c, s) = |I(c) - I(s)| \quad (1)$$

$$E(c, s) = |E(c) - E(s)| \quad (2)$$

$$Sym(c,s)=|Sym(c) - Sym(s)| \quad (3)$$

$$RG(c,s)=|R(c) - G(c)| - |G(s) - R(s)| \quad (4)$$

$$BY(c,s)=|B(c) - Y(c)| - |Y(s) - B(s)| \quad (5)$$

where "-" represents interpolation to the finer scale and point-by-point subtraction. Totally, 30 feature bases are computed because the five center-surround feature basis individually have 6 different scales [6, 9]. The 30 feature bases with variant scale information are combined into four feature maps as shown in Eq. (6) where $\overline{I}$, $\overline{E}$, $\overline{Sym}$ and $\overline{C}$ stand for intensity, edge, symmetry, and color opponency, respectively. These are obtained through across-scale addition "$\oplus$" [6, 9].

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)), \quad \overline{E} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(E(c,s)),$$

$$\overline{Sym}= \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(S(c,s)), \overline{C}= \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(RG(c,s)+BY(c,s)) \quad (6)$$

2.3 ICA algorithm

In this article, we use unsupervised learning to determine the relative importance of different bases used to generate a suitable salient region using ICA. Even though it is difficult to understand the mechanisms of the human brain including the visual cortex to process the complex natural scene, Barlow's hypothesis might be useful in explaining the role of the human brain.

We suppose that eye movements in bottom-up processing are the result of our brain activity for maximizing visual information, and that the eye sequence that we focus on an object can be modeled by redundancy reduction of the visual information. In order to model the saliency map, we use ICA because the ICA algorithm is the best way to reduce redundancy [7]. The ICA algorithm is able to separate the original independent signals from the mixed signal by learning the weights of the neural network used to maximize the entropy or log-likelihood of output signals. Additionally, the ICA algorithm can extract important features to minimize the redundancy or mutual information between output signals [7-8]. The purpose of ICA

is to seek mutually independent components, and it leads to a local representation quite similar to that obtained through sparse coding [8].

We consider intensity, edge, and color opponent coding in the retina and symmetry in the LGN, and use these results for input patches of ICA. **Figure 4** shows the procedure of realizing the saliency map from four feature maps, $\overline{I}$, $\overline{E}$, $\overline{Sym}$ and $\overline{C}$. In **Figure 4**, Eri is obtained by the convolution between the r-th channel of the feature maps(FMr) and the i-th filters(ICsri) obtained by ICA learning as shown in Eq. (7):

$$E_{ri} = FM_r * ICs_{ri} \quad for\ i = 1,...,N, \quad r = 1,..,4 \tag{7}$$

where N denotes the number of filters. Convoluted feature map, Eri represents the influences of the four feature maps have on each independent component. Finally, a saliency map is obtained using Eq. (8):

$$S(x, y) = \sum E_{ri}(x, y) \quad for \quad all \quad i \tag{8}$$

The saliency map S(x,y)is computed by summation of all feature maps for every location (x, y) in an input image. A salient location P is the maximum summation value in a specific window of a saliency map, as shown in Eq. (9):

$$\boldsymbol{P} = \arg\max_{(x,y)} \left\{ \sum_{(u,v)\in W} S(u,v) \quad for \quad all \quad (x,y) \right\} \tag{9}$$

where (u,v) is a window with 20×20 size. The selected salient location P is the most salient location of an input image.

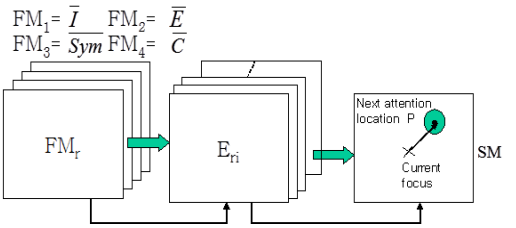


Figure 4: Realizing of the saliency map model from feature maps

($\overline{I}$: intensity feature map, $\overline{E}$: edge feature map, $\overline{Sym}$: symmetry feature map, $\overline{C}$: color feature map, SM : saliency map)

In our brain, there is an inhibition-of-return (IOR) function [11]. In order to implement the IOR function, we use a symmetry information on the object when the most salient region contains an object because the object has commonly symmetrical property. The IOR region is obtained by the noise-tolerant generalized symmetry transformation (NTGST) and the dilation function of the morphology algorithm [9]. When the salient region doesn't include an object, we consider the IOR region that has a high pixel value above the specified threshold in the saliency map. After masking this region, the SM finds the next salient point that excludes the previous salient object [10]. Comparing the maximum salient values in two camera images, we decide the master eye that has a camera with larger salient value. In order to verify the candidate as a landmark, we need to compare the salient region of the master eye with that of the slave eye. The regions obtained by IOR function, which is to avoid the duplicate selection of the most salient region, are compared to decide the landmark. If the IOR region of the master eye is similar to that of the slave eye, we regard the IOR region as a landmark to make convergence. The comparison of histogram values of IOR regions between left and right cameras are used for the verification of a landmark. After a landmark selection is successfully made, the depth information is calculated. **Figure 5** shows the top view of verged cameras.

2.4 3D detecting Algorithm

This scheme is applied to the unmanned vehicle so that the unmanned vehicle can decide a direction and calculate the distance to move to the target.

The 3 D detecting algoritm is described in reference [4].

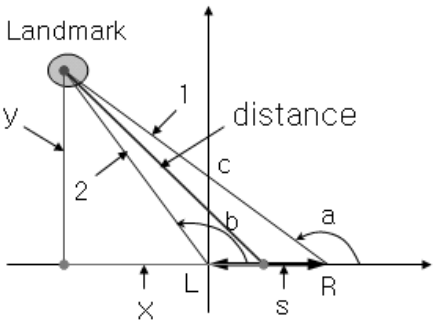


Figure 5: Top view of verged cameras

In **Figure 5**, the parameters related with the distance are obtained by mathematical relationship, and they are defined as

L : Left camera focus center, R : Right camera focus center, a : Right camera angle, b : Left camera angle, c : an intercept of the line 1, s : The distance between two cameras focus, 1 and 2 : the straight line from right and left cameras to a landmark

Considering the limitation of the field view (F) in horizontal axis and motor encoder resolution (U), we can get the encoder value (E) for the limited field view of horizontal axis. Eq. (10) shows the encoder value (E). This encoder value is used to calculate the encoder value (xt) of the horizontal axis motor for aligning each camera to a landmark as shown in Eq. (10). In Eq. (11), R denotes the x-axis resolution of the image and T denotes the x coordinates of a landmark. The encoder value to move each camera to the landmark point is translated to the angel value(xd) by Eq. (11). Resultantly, the angles a and b are obtained by Eq. (12) by substitution T into the x coordinates of the left and right cameras. We don't need to consider the y coordinates for which the values of IOR regions of two cameras become almost same.

$$E = (F*360°)/U \tag{10}$$

$$xt = - E +(E*T)/R \tag{11}$$

$$xd = 90° - (R*xt)/U \tag{12}$$

$$\tan(a)\cdot x - s\cdot\tan(a) = y \tag{13}$$

$$\tan(\quad b\ )\cdot x = y \tag{14}$$

$$y = \frac{\tan(\quad a\ )\cdot\tan(\quad b\ )}{\tan(\quad a\ )-\tan(\quad b\ )}\cdot s \tag{15}$$

Eq. (13) and (14) show the equation of straight lines between the cameras and the landmark, respectively. Eq. (14) is the equation to get the vertical distance (y) in **Figure 5**.

If the angles of a and b shown in **Figure 5** are above 90° (case 1) in **Figure 6**, the distance is $\sqrt{x'^2+y^2}$ because of $x = y/\tan(b)$ and $x' = |y/\tan(b)| + s/2$. If the angle a is above 90° and the angle b is less than 90° (case 2), the distance is almost y because the error is very small. If the angles of a and b are under 90 (case 3), the distance is $\sqrt{x''^2+y^2}$ because of $x = y/\tan(b)$ and $x'' = |y/\tan(b)| - s/2$.

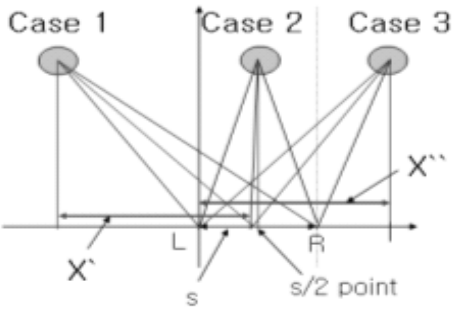


Figure 6: Three cases of obtaining the depth information

Figure 7 shows a block diagram of the implemented active vision system and active stereo

vision system. We use two CCD type cameras as the image sensor, and two image data are obtained by MIL image grabber and transferred to the IBM PC at a speed of 30 frames per second. The saliency map model which is implemented in IBM PC respectively generates a target point and transfers it to the DSP board using RS232C serial communication at each sampling instant. The PID controller with suitable control gains generates a proper motor control signal that is used to bring the respective selected salient area into the focus of CCD camera.

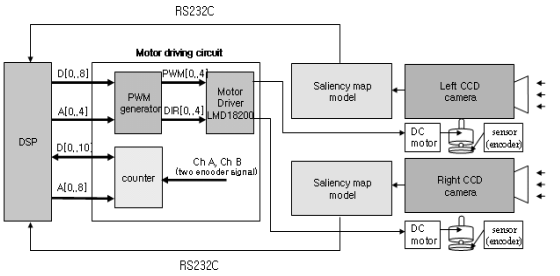


Figure 7: The hardware setup for active stereo vision system

3. Experiment of the active stereo vision system

3.1 Composition of the unmanned vehicle with the stereo vision system

To implement the developed active stereo vision system, a prototype of an unmanned vehicle was built as shown in **Figure 8**. The vehicle has a track for forward motion and orientation change actuated by two DC motors. Two 150W Maxon motor with developed DC motor drivers and I/O(input-output) interface system were used to the vehicle. The vehicle was designed to be autonomous such that it is boarded with the stereo vision system, controller, driver, and batteries. As shown in **Figure 9**, the structure of the developed control system for the motion of the leg system is composed of the DC servo motor driver, motion controller, main control system, and I/O interface system. The sequential motion of the unmanned vehicle is controlled by the PID control loop, and the whole moving motion is intelligently controlled using the vision and encoder sensor feedback.

Figure 8: Outlook of the unmanned vehicle with the stereo vision system

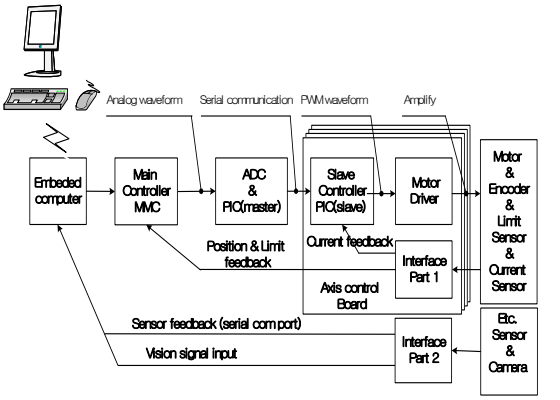
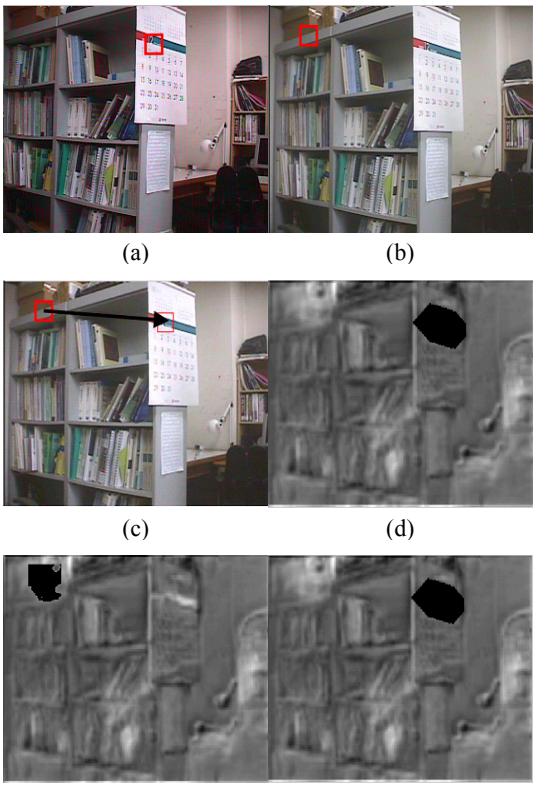


Figure 9: The whole control structure for the stereo vision system

3.2 Experiment of the active stereo vision system

Figure 10 (a) and (b) show a salient region of a master and a slave eye, respectively. **Figure 10** (d) and (e) show the SM with the IOR functions. As shown in these figures, the y coordinates of the salient regions for the master and the slave eye are almost the same, but the comparison of the IOR regions gives a different landmark. In this case, we consider the next salient region in the same y coordinate, in which the IOR function prevents the

duplicate selection of most salient region. **Figure 10** (c) and (f) show the second salient region of the slave eye and the SM with the IOR function, respectively. Comparing **Figure 10**(d) with **Figure 10**(f), the IOR region is almost the same, as a result we can obtain a landmark for convergence. **Figure 11** shows the verification of a landmark by comparing the IOR regions in occlusion region. Also, **Figure 12** shows the vergence control results of the proposed active stereo vision system using the SM model.

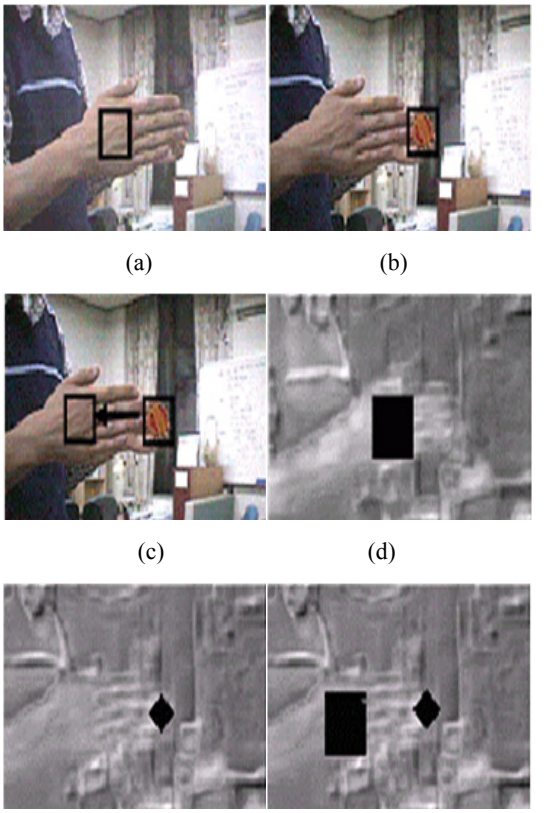


(a) (b)

(c) (d)

(e) (f)

Figure 10: Verification of a landmark by comparing the IOR regions (a) most salient region of the master eye, (b) most salient region of the slave eye, (c) the second salient region of the slave eye (d) IOR region of most salient region in the master eye, (e) IOR region of most salient region in the slave eye, (f) IOR region of the second salient region in the slave eye



(a) (b)

(c) (d)

(e) (f)

Figure 11: The verification of a landmark by comparing the IOR regions in occlusion region

(a) The most salient region of the master camera, (b) The most salient region of the slave camera, (c) The second salient region of the slave camera, (d) The IOR region of most salient region in the master camera, (e) The IOR region of occlusion region and the IOR region of the most salient region in the slave camera, (f) The IOR region of the second salient region in the master camera

After finding out the depth through vergence control results of active vision stereo system, the depth information is sent to the unmaned vehicle for moving to the target.

(a)

(b)

(c)

Figure 12: Vergence control results of active vision stereo system using the SM (a) The first landmark selection of the left and the right cameras (b) The first SM of the left and the right cameras (c) The first result of vergence control by motor control

4. Conclusion

An intelligent stereo vision sensor system is developed in this paper. The active stereo vision system that mimics human-like bottom-up visual attention to identify the direction and depth. The human-like bottom-up visual attention is realized by implementing a saliency map model that reflects the biological visual pathway from retina cells to visual cortex through the LGN. Through the experimental results of the vision system, the effectiveness of the implemented active stereo vision system is verified.

Further experimental research will be undertaken applying the developed stereo vision system to the unmanned vehicle such as unmanned forklift in near future.

Acknowledgment

This research was supported by a grant (11Transportation System-Logistics02) from Transportation System Efficiency Program funded by Ministry of Land Transport and Maritime Affairs of Korean government

References

[1] M. Seelinger and J. D. Yoder, "Automatic visual guidance of a forklift engaging a pallet," Robotics and Autonomous Systems, vol. 54, pp. 1026-1038, 2006.

[2] S. M. Byun and M. H. Kim, "Pallet measurement Method for Auomatic Pallet Engaging in Real time", Journal of Korea Multimedia Society, vol. 14, no. 2., pp. 171-181, 2011.

[3] A. Bernardino and J. Santos-Victor, "Vergence control for robotic heads using log-polar images", Intelligent Robots and Systems, vol. 3, pp. 1264-1271, 1996.

[4] J. Batista, P. Peixoto, H. Araujo, "A focusing-by-vergence system controlled by retinal motion disparity", Proceedings of IEEE International Conference Robotics and Automation, vol. 4, pp. 3209-3214, 2001.

[5] J. Conradt, M. Pescatore, S. Pascal, and P. Verschure, "Saliency Maps Operating on Stereo Images Detect Landmarks and their Distance", Lecture Notesin Computer Science. pp. 347-358, 2002.

[6] S. J. Park, J. K. Shin, M. Lee, "Biologically inspired saliency map model for bottom-up visual attention", Lecture Notesin Computer Science. pp. 418-426, 2002.

[7] A. J. Bell and T. J. Sejnowski, "The independent components of natural scenes are

edge filters", Vision Research, vol. 37, pp. 327-333, 1997.

[8] T. W. Lee, Independent Component Analysis-Theory and Application, Kluwer Academic Publisher, 1998.

[9] C. J. Park, W. G. Oh, S. H. Cho, and H. M. Choi, "An efficient context-free attention operator for BLU inspection of LCD", Proceedings of, pp. 251-256, 2000.

[10] 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.

[11] L. Itti and C. Koch, "Computational Modeling of Visual Attention", Nature Reviews Neuroscience, vol. 2, pp. 194-203, 2001.