# A Joint Learning Framework of Visual Sensory Representation, Eye Movements and Depth Representation For Developmental Robotic Agents

Tanapol Prucksakorn, Sungmoon Jeong*, and Nak Young Chong

School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, Japan. Email: tanapol.pr@jaist.ac.jp, jeongsm@jaist.ac.jp, nakyoung@jaist.ac.jp

**Abstract.** In this paper, we propose a novel visual learning framework for developmental robotics agents which mimics the developmental learning concept from human infants. It can be applied to an agent to autonomously perceive depths by simultaneously developing its visual sensory representation, eye movement control, and depth representation knowledge through integrating multiple visual depth cues during self-induced lateral body movement. Based on the active efficient coding theory (AEC), a sparse coding and a reinforcement learning are tightly coupled with each other by sharing a unify cost function to update the performance of the sensory coding model and eye motor control. The generated multiple eye motor control signals for different visual depth cues are used together as inputs for the multi-layer neural networks for representing the given depth from simple human-robot interaction. We have shown that the proposed learning framework, which is implemented on the Hoap-3 humanoid robot simulator, can effectively learn to autonomously develop the sensory visual representation, eye motor control, and depth perception with self-calibrating ability at the same time.

## 1 Introduction

For living organisms such as humans and mammals, visual perception is one of the most important function. It gives the organism an ability to learn and interact with environments around them. However, when they were born, they do not instantly understand how to use the information they perceived. So, for their lifetime they continuously learn and improve their perception, while interact with the environments. In biological vision systems, the data that is collected by human or animals organs are very noisy and messy data. It is not self-explanatory meaningful information [10]. So, it is quite difficult for us to make use of these non-obvious data. In [17], they discussed that our brain is not programmed to know how to use those data, but instead the brain is trained

* The first two authors contributed equally.

autonomously to understand how to translate those noisy unordered information into visual perception. In the same way, the developmental robotics agents which are not programmed with visual perception ability faced the same problem that they do not know how to utilize the data. Thus, to use those information, we must create a representation of the data that is packed with the vast information.

In the previous studies, an active efficient coding proposed in [1, 2, 4] is employed to encode and represent the information perceived by the robot by taking advantages of redundancies. Reinforcement learning algorithm is used as a learning scheme for the robot to generate eye movements based on the encoded information. It has been proven to be successful when learning of vergence and smooth pursuit eye movement are needed [7, 12, 15, 18, 19]. In [16], they have successfully demonstrated generating multiple eye movements, which are smooth pursuit and vergence to track a moving object, but depth perception is not included in the learning framework. Moreover, all of the generated eye movement information could not be used for depth perception because stationary observer cannot extract depth information from motion parallax or optic flow without a priori knowledge such as object size.

Especially, to actively perceive the depth information, the biological vision systems can autonomously generate multiple visual depth cues during the lateral body movement such as stereo disparity and motion parallax. When they keep the visual fixation during the body movement, both of the visual depth cues and eye movements are autonomously generated by the same intrinsically motivated learning principal to maximize the redundancy between sensory inputs in binocular and monocular viewing. In [12], they have proposed a developmental learning framework for active depth perception with self-induced lateral body movement, but they only considered a single depth cue as motion parallax. Interestingly, the organism does not use only one visual depth cue for their whole lifetime. They can integrate the information about multiple visual depth cues and analyze the eye movements to perceive the spatial information about the surrounding environment. Generally, in psychology, dominant eye is a concept that implies that one eye moves before another eye does. Recently there are studies that support the dominant eye hypothesis [6, 13, 14]. Also, according to [5], they reported that when a motion is self-induced by active observer, two visual depth cues (stereo vision and motion parallax) will be sequentially activated which is not observable in a static observer. Therefore, we may consider that two eye movements for different visual depth cues during the self-induced lateral body movement can be sequentially generated in an independent process to minimize the conflict of multiple cues and then finally multiple eye movements are used to analyze the depth information by integrating each of them. To the best of our knowledge, no one has attempted to propose such a learning framework for developmental robots under the efficient coding theory. This approach enables to autonomously learn not only sensory representation and eye movement controls for the multiple visual depth cue analysis but also active depth perception during self-induced body movements.

## 2 Methods

We combine the sparse coding and reinforcement learning algorithm together to achieve active efficient coding for learning of multiple cues from the dominant and non-dominant eyes, respectively. Sensory coding model learns how to encode and represent the two images which are generated by the dominant eye with self-induced lateral body movement for motion parallax and two eyes for stereo disparity. Reinforcement learner controls the motor based on the encoded information done by the sensory coding model to increase the efficiency of the coding model.

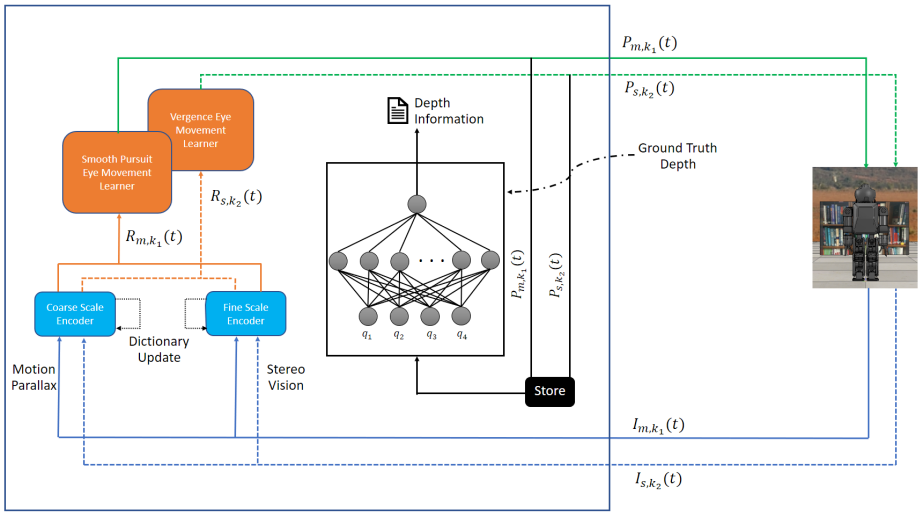



**Fig. 1.** Model architecture. (1) At the first step $k_1$, to perform the motion parallax, the robot captures the successive images $I_{m,k_1}(t)$ during the self-induced lateral body movement which are fed into the sensory encoders with multiple image scales. Later, an output reward signal, $R_{m,k_1}(t)$, is sent to the reinforcement learner to generate an appropriate eye movement to hold the fixation during the body movement. Finally, pan command $P_{m,k_1}(t)$ is sent to the robot and it generates the smooth pursuit eye movement for dominant eye camera to maximize the redundancy between the successive images. (2) At the second step $k_2$, stereo images $I_{s,k_2}(t)$ are captured from both two cameras and sent to the sensory encoders. An output reward signal, $R_{s,k_2}(t)$, is sent to the reinforcement learner to generate the vergence command $P_{s,k_2}(t)$ to maximize the redundancy between the stereo images. The visual dictionaries are then updated based on visual reconstruction errors for both of visual depth cues. Finally, the stored eye movements ($q_1$, $q_2$, $q_3$, and $q_4$) are used as an input for the neural network to represent the depth information which is given by human-robot interaction.

### 2.1 Model Architectures

Since the concept of integrating two cues with dominant eye requires that one eye should move before the another one, the framework is divided into 2 steps which are motion parallax for the dominant eye first and stereo vision for the non-dominant eye. The framework (Fig. 1) sequentially perform motion parallax, stereo vision and their integration to represent the depth information. In one iteration $t$, it is subdivided to 2 steps $k_1$ and $k_2$. At step $k_1$, the robot will perform motion parallax by moving laterally from original position to the *leftmost position*. Then at sub-iteration $k_2$, the robot will execute stereo vision after the motion parallax is done. After $h$ iterations, the robot will perform the motion parallax again, but it will move laterally to the *rightmost position*. Then, the stereo vision is performed and the entire process is repeated for another $h$ iterations with a certain visual fixation with different texture of the object and depth between the robot and the object.

Step $k_1$: the framework receives the input image from the dominant eye as the monocular viewing. Two successive images $I_{m_1}(t)$ and $I_{m_2}(t)$ are captured at different positions during self-induced lateral body movement. The two images $I_{m,k_1}(t) = \left[I_{m_1}(t)\ I_{m_2}(t)\right]$ are then used as an input for the framework to learn not only sensory representation of motion parallax but also smooth pursuit eye movement for the dominant eye.

Step $k_2$: after the smooth pursuit eye movement learner successfully sent the pan command, the dominant eye panned respect to the command. Then both eyes capture images $I_{s_1}(t)$ and $I_{s_2}(t)$ which are combined to $I_{s,k_2}(t) = \left[I_{s_1}(t)\ I_{s_2}(t)\right]$. The stereo images are sent to the framework to learn the sensory representation of stereo disparity and vergence eye movement for the non-dominant eye.

### 2.2 Sensory Coding Model

Two input images are then cropped by 128x128 pixels and 80x80 pixels from the center of the images. Two cropped images represent fine scale and coarse scale respectively. We use two scales of the images to represent the foveal system in human eyes. The fine scale image represents a foveal region ion our eyes which has more detail from the center of vision. While, coarse scale represents parafoveal area which has lower detail. Discussions and comparisons between using one scale and two scales have been done in [7]. They discussed how gaining the access of multi-scale images could improve the learning of the framework. While, having only one scale might prevent the system to learn.

After cropping, the cropped images are then convert to gray scale. 10 by 10 pixels patches are extracted from the gray scale images whose locations are generated by 1 pixel and 4 pixels shifts horizontally and vertically for coarse scale and fine scale, respectively. The image patches are then sub-sampled using Gaussian pyramid algorithm by a factor of 8 for coarse scale, and factor of 2 for fine scale. The patches are reshaped to be one-dimensional vectors which

have zero mean and unit norm, $x_i^j(t)$. Where, $i$ is the index of the patch, and $j \in \{C, F\}$. $C$ is for coarse scale and $F$ stands for fine scale.

For coarse scale and fine scale, the two one-dimensional vectors are then combined into a single vector $x^j(t)$. The first 100 elements of the vectors are from the first image and the remaining are from the second image. The combined vectors ($x^C(t)$ and $x^F(t)$) will consist of $P = 200$ elements.

Later, the patches are encoded by sparse coding algorithm in linear fashion. Each patch can be represented by a linear combination of basis functions picked from an over-complete dictionary $\phi^j(t) = \{\phi_n^j(t)\}_{n=1}^N$ [11]. We use $N = 288$ basis functions. Two pairs of dictionaries are randomly initialized and normalized each pair contains coarse scale and fine scale dictionary for stereo vision ($d = s$) and motion parallax ($d = m$) as shown in Fig. 1. We use matching pursuit algorithm [8] to estimate and find the sparse representation of the input vector by the weighted sum as follows:

$$x_i^j(t) \approx \hat{x}_i^j(t) = \sum_{n=1}^{N} a_{i,n}^j(t)\phi_n^j(t) \ . \tag{1}$$

The matching pursuit algorithm suits to concept of sparse coding, which can estimate $x_i(t)$ by using a limited number of coefficients. In this research, the maximum number of non-zero scalar coefficients $a_{i,n}(t)$ is set to be 10 elements to ensure sparseness of the efficient coding. For later use in reinforcement learner part, pooled activity, $f_n(t)$, which represent the activity of each neuron cell is calculated from the coefficients from matching pursuit algorithm as follows:

$$f^j(t) = \begin{bmatrix} f_1^j(t) \\ f_2^j(t) \\ \vdots \\ f_P^j(t) \end{bmatrix} \ . \tag{2}$$

Where, each element of the vector $f^j(t)$ is described as:

$$f_n(t) = \sum_{i=1}^{P} a_{i,n}(t)^2. \tag{3}$$

A reconstruction error is introduced as a cost function to be used in sensory coding model and reinforcement learner. It measures the estimation error of vector $x(t)$. The reconstruction error is defined as:

$$e(t) = \frac{1}{P} \sum_{i=1}^{P} \frac{\|x_i(t) - \sum_{n=1}^{N} a_{i,n}(t)\phi_n(t)\|^2}{\|x_i(t)^2\|}. \tag{4}$$

Gradient descent method is used to update the dictionaries with the reconstruction error as a cost function. After each update, the dictionaries are then normalized.

### 2.3 Reinforcement Learning

The state representation of the reinforcement learner can be described by combination of coarse scale and fine scale pooled activity, $f_n(t)$ as follows:

$$f(t) = \begin{bmatrix} f^C(t) \\ f^F(t) \end{bmatrix} . \tag{5}$$

The reward that is given to the learning agent is a negative of the summation of reconstruction error from both scales which is described as:

$$R_{d,k}(t) = -(e^C(t) + e^F(t)) . \tag{6}$$

Where, $k \in \{k_1, k_2\}$ and $d \in \{m, s\}$. $m$ is for motion parallax. $s$ is for stereo vision. An actor-critic algorithm number 3 proposed in [3] is employed for the leaner agent. For action selection, we use Gibbs distribution (softmax) for probabilistically choosing an action as follows:

$$\pi(f(t), a_t) = \frac{e^{z_a}}{\sum_{a' \in A} e^{z_{a'}}} . \tag{7}$$

For each action, the activation value $z_a$ is given by:

$$z_a = \sum_{n=1}^{N} w_a(t) f_n(t) , \tag{8}$$

where $w_a(t)$ is a weight vector from the state $f(t)$ to action $a$ that is initially random. The action is pan angle of the cameras in degrees. Possible actions $a$ are contained in a set of actions $A$. In this research we use $A = \{-0.2°, -0.1°, -0.05°, 0°, 0.05°, 0.1°, 0.2°\}$. Thus, the policy maps $f(t)$ to $a \in A$. The selected actions are $P_{m,k_1}(t)$ for motion parallax and $P_{s,k_2}(t)$ for stereo vision.

### 2.4 Depth Representation

A simple feed forward neural network with two layer is used to interpret between eye movements to the object's distance. In each iteration after stereo vision is executed, the eye movements are stored and accumulated for depth estimation. When the robot successfully performs motion parallax and stereo vision at both leftmost position and rightmost position, the amount of eye movements $\boldsymbol{q}$ are then used to train the neural network. $\boldsymbol{q}$ contains:

1. $q_1$, Left eye's pan movement at leftmost position
2. $q_2$, Vergence eye movement at leftmost position
3. $q_3$, Left eye's pan movement at rightmost position
4. $q_4$, Vergence eye movement at rightmost position

We use Levengerg-Marquardt method [9] for training the neural network. A sigmoid transfer function is used in the hidden layer which has 10 neurons. The input of the neural network is $\boldsymbol{q}$. While, the target is ground truth depth provided by supervisor.

## 3 Simulations & Results

### 3.1 Experimental Setup

We use V-REP, a robot simulator, as a 3D environment visualization for the framework. The framework is implemented and developed in MATLAB. The environment in the simulator comprises HOAP3 robot, an object with interchangeable texture, and a still background image. The lateral movement of the robot is simplified to be changing the position of the robot directly to cut out the travel time.

In this simulation, we test the multiple cues to estimate the depth between the robot and the object and it is from 1 meter to 3 meters with an 0.1 meter interval, i.e. 1.0,1.1,1.2,...3.0 meters. The distance between the leftmost position and the rightmost position is 0.2 meter, i.e. $\delta = 0.1$. The baseline, distance between two eyes, is 0.06 meter. The number of iterations $h$ is 30 iterations. We prepare 100 different images to learn the various visual textures of the environment. To evaluate the eye movement training, we define mean absolute error (MAE) for evaluating eye movements as follows:

$$\mathrm{MAE(t)} = \frac{1}{1000}\sum_{\mathrm{k}=0}^{999} \left|\theta(\mathrm{t}+29+30\mathrm{k}) - \theta^*(\mathrm{t}+29+30\mathrm{k})\right| . \tag{9}$$

Where,

1. $\theta(t)$ represents the pan/vergence angle of the eye at time $t$
2. $\theta^*(t)$ represents the optimal pan/vergence angle at time $t$

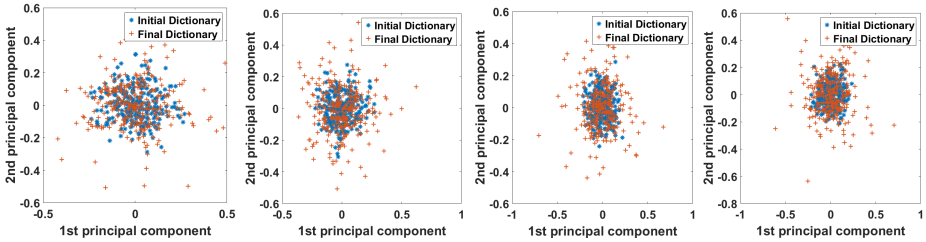
### 3.2 Adaptive Visual Dictionary

The principal component analysis (PCA) is applied to visualize the distribution of the visual dictionaries. Because the visual dictionaries were randomly initialized, most of the elements are quite redundant between each other. The first and the second PCs are used to visualize the distribution of visual dictionary as shown in Figs. 2(a) - 2(c). We can see that the trained visual dictionaries are more sparsely distributed than the initial dictionary.

### 3.3 Joint Development of Active Depth Perception

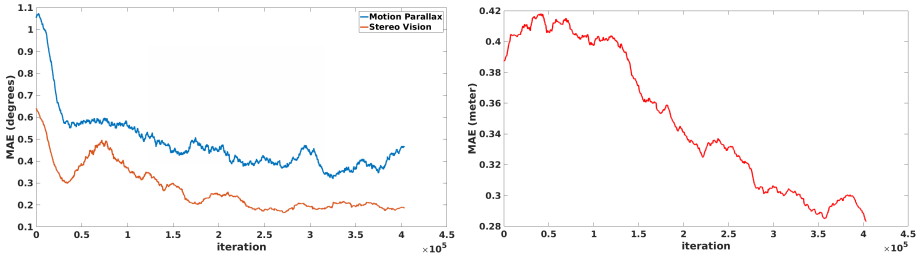
The results of the training are shown in Fig. 3. Fig. 3(a) shows the MAE of eye movements. The red line shows the MAE of the stereo vision, while the blue line shows the MAE of the smooth pursuit. To test the depth perception, all of the eye movements $\boldsymbol{q}$ with different experimental conditions are used as inputs for the neural networks. The outputs from the neural network are used to calculate the MAE at every time steps. We applied a moving average window with window size of 1,000 iterations to observe trend of the depth learning as shown in Fig. 3(b).

From the simulation results, we can see that the framework could jointly learn to improve the sensory encoding and represent the visual stimuli while learning to generate multiple eye motor control with depth perception.




(a) Coarse scale for stereo disparity (b) Fine scale for stereo disparity (c) Coarse scale for motion parallax (d) Fine scale for motion parallax

**Fig. 2.** Visualization of development of the visual dictionaries. The distribution of the visual dictionaries using the first and second PCs at the initial time and the end of training, respectively.

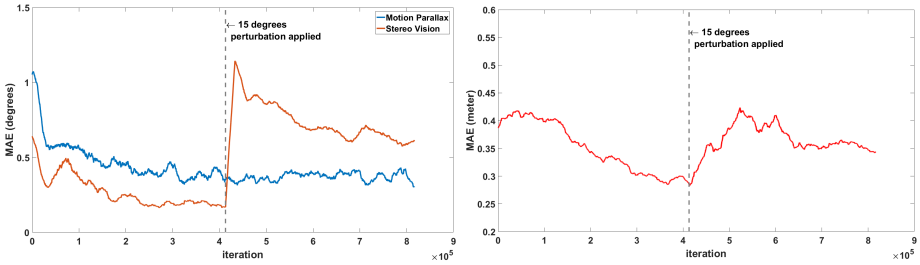



(a) MAE of eye movements (b) MAE of depth perception

**Fig. 3.** Development in each part of the system. The figures visualize the evolution of the visual representation (coding), eye movement and depth estimation. (a) represents the eye movement errors in form of mean absolute error. (b) shows how depth estimation develops through the learning.

## 4 Robustness Test

To verify the adaptation properties of the framework, perturbations are simulated by applying a constant in-plane roll rotation of each camera at a time as shown in Fig. 4. In Fig. 4(a), noticeable increases in eye movement errors are observed after inducing the disturbance which are presented by the gray dotted line in the figures. Smooth pursuit eye movements are not largely effected by the disturbance, while vergence eye movements are more susceptible to the interference. Because the vergence eye movements are dependent to the results of smooth pursuit eye movements. Even though the vergence control MAE is drastically increased as shown in Fig. 4(a), the MAE of depth perception is slowly increased at that time. Because, the depth perception is done by integrating both of eye movements and it could be recovered with the supports from both cues as shown in Fig. 4(b).




(a) MAE of eye movements with perturbation

(b) MAE of depth perception with perturbation

**Fig. 4.** Adaptation property from the perturbation. (a) MAE of the eye movements during execution of learning time with the perturbation. (b) MAE of the depth perception during execution of learning time with the perturbation

## 5 Conclusion

In this research, we proposed a novel developmental learning framework to actively the active depth perception during self-induced lateral body movements. The proposed framework can simultaneously develop the sensory representation, eye movement control and integration of the visual depth cues such as stereo disparity and motion parallax. In order to avoid the conflict of multiple eye movements, the two different eye movements are sequentially trained and generated, while they share the same learning architecture. Finally, the generated multiple eye movements are effectively used to represent the depth information. Also, the proposed learning framework can be seamlessly recovered from the external perturbations. To extend this to fully autonomous architecture, an unsupervised learning method will be employed instead. Moreover, the dominant eye may be competitively selected during the learning period.

**Acknowledgement** This work was supported by Japan-Germany collaboration research project on computational neuroscience "Autonomous Learning of Active Depth Perception: from Neural Models to Humanoid Robots" from Japan Agency for Medical Research and Development (AMED) and was partially supported by EU-Japan coordinated R&D project on "Culture Aware Robots and Environmental Sensor Systems for Elderly Support" commissioned by the Ministry of Internal Affairs and Communications (MIC) of Japan and EC Horizon 2020.

## References

1. Attneave, F.: Some informational aspects of visual perception. Psychol. Rev pp. 183–193 (1954)

2. Barlow, H.B.: Possible principles underlying the transformation of sensory messages. Cambridge, MA: MIT Press (1961)
3. Bhatnagar, S., Sutton, R.S., Ghavamzadeh, M., Lee, M.: Natural actor–critic algorithms. Automatica 45(11), 2471–2482 (2009)
4. Field, D.J.: What is the goal of sensory coding? Neural Comput. 6(4), 559–601 (Jul 1994)
5. Frey, J., Ringach, D.L.: Binocular eye movements evoked by self-induced motion parallax. The Journal of Neuroscience 31(47), 17069–17073 (2011)
6. Johansson, J., Seimyr, G.Ö., Pansell, T.: Eye dominance in binocular viewing conditions. Journal of vision 15(9), 21–21 (2015)
7. Lonini, L., Zhao, Y., Chandrashekhariah, P., Shi, B., Triesch, J.: Autonomous learning of active multi-scale binocular vision. In: Development and Learning and Epigenetic Robotics (ICDL), 2013 IEEE Third Joint International Conference on. pp. 1–6 (Aug 2013)
8. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. Signal Processing, IEEE Transactions on 41(12), 3397–3415 (1993)
9. Moré, J.J.: The levenberg-marquardt algorithm: implementation and theory. In: Numerical analysis, pp. 105–116. Springer (1978)
10. Mugan, J., Kuipers, B.: Autonomous learning of high-level states and actions in continuous environments. Autonomous Mental Development, IEEE Transactions on 4(1), 70–86 (March 2012)
11. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: A strategy employed by v1? Vision Research 37(23), 3311 – 3325 (1997)
12. Prucksakorn, T., Jeong, S., Triesch, J., Lee, H., Chong, N.Y.: Self-calibrating active depth perception via motion parallax. In: Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2016 Joint IEEE International Conferences on. IEEE (2016)
13. Shneor, E., Hochstein, S.: Eye dominance effects in feature search. Vision research 46(25), 4258–4269 (2006)
14. Shneor, E., Hochstein, S.: Eye dominance effects in conjunction search. Vision research 48(15), 1592–1602 (2008)
15. Teulière, C., Forestier, S., Lonini, L., Zhang, C., Zhao, Y., Shi, B., Triesch, J.: Self-calibrating smooth pursuit through active efficient coding. Robotics and Autonomous Systems 71, 3–12 (2015)
16. Vikram, T., Teuliere, C., Zhang, C., Shi, B., Triesch, J.: Autonomous learning of smooth pursuit and vergence through active efficient coding. In: Development and Learning and Epigenetic Robotics (ICDL-Epirob), 2014 Joint IEEE International Conferences on. pp. 448–453. IEEE (2014)
17. Weng, J., Luciw, M.: Brain-like emergent spatial processing. Autonomous Mental Development, IEEE Transactions on 4(2), 161–185 (June 2012)
18. Zhang, C., Zhao, Y., Triesch, J., Shi, B.E.: Intrinsically motivated learning of visual motion perception and smooth pursuit. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on. pp. 1902–1908. IEEE (2014)
19. Zhao, Y., Rothkopf, C., Triesch, J., Shi, B.: A unified model of the joint development of disparity selectivity and vergence control. In: Development and Learning and Epigenetic Robotics (ICDL), 2012 IEEE International Conference on. pp. 1–6 (Nov 2012)