Perceptual Attention Through Image Feature Generation Based on Visio-Motor Map Learning

Minoru Asada and Takashi Minato
Dept. of Adaptive Machine Systems
Graduate School of Engineering
Osaka University
Suita, Osaka 565-0871, Japan
e-mail: asada@ams.eng.osaka-u.ac.jp, minato@er.ams.eng.osaka-u.ac.jp

Abstract

Perceptual attention can be regraded as the first step towards symbol emergence from senroy data. Especially, visual attention is one of the key issues for robots to accomplish the given tasks, and the existing methods specify the image features and attention control scheme in advance according to the task and the robot. However, in order to cope with environmental changes and/or task variations, the robot should construct its own attention mechanism. This paper presents a method for image feature generation by visio-motor map learning for a mobile robot. The teaching data constructs the visio-motor mapping that constrains the image feature generation and state vector estimation as well. The resultant projection matrix from the filtered image to a state vector tells us which part of the image is more informative for decision making than others. The method is applied to indoor navigation and soccer shooting tasks, and discussion is given.

Introduction

Through billions of years of evolution process, biological systems have acquired their organs and their own strategies so that they can survive in hostile environments. Visual attention can be regarded as a combination of such organs and strategies, that is, "vision" that brings a huge amount of data about the external world and "attention mechanism" that extracts the necessary and sufficient information from them for the system to achieve the mission at hand. Such a capability is desired in artificial systems too, and therefore, it has been one of the most typical but formidable issues in robotics and AI for long years.

Human beings can easily enjoy such a mechanism in various kinds of situations, and a number of researches focus on the early visual processing of human beings (Treisman & Gelade 1980), improve the Treisman' model (Wolfe, Cave, & Franzel 1989; Laar & Gielen 1997), and apply Shanon's information of the observed image (Takeuchi, Ohnishi, & Sugie 1998) in order to select the focus of attention in the view. The main issues of these works are the analysis of the human visual processing and the explanation for our attention mechanism.

Some of computer vision researchers focused on the view point selection (where to look) problem (Nayar, Murase, & Nene 1996; Arbel & Ferrie 1999) in order to disambiguate the descriptions for the observed image that is obtained by matching the image with the model database. The selection criterion is based on the statistics of the image data and actions (gaze control), if any, are intended to get the better observation for object recognition, but are not directly related to physical actions needed to accomplish a given task.

Self localization is the one of the issues in navigation task, and most of the works are based on a kind of geometric reconstruction from the observed image using a priori knowledge of the environment. Thrun (Thrun 1998) and Vlassis et al. (Vlassis, Bunschoten, & Kröse 2001) extracted the features correlated to the information of the self-localization of the mobile robot from the observed images based on the probabilistic method. Kröse and Bunschoten (Kröse & Bunschoten 1999) decided the robot direction, i.e., camera direction by minimizing the conditional entropy of the robot position given the observations.

The existing approaches mentioned above mostly specify the kinds of image features in advance and adopt a sort of attention mechanism based on the designers intuition having considered the given task. However, in order to cope with environmental changes and/or task variations, the robot should generate image features and construct its own attention mechanism. This paper presents a method for image feature generation by visio-motor map learning for a mobile robot. The teaching data construct the visio-motor mapping that constrains the image feature generation and state vector estimation for the action selection as well. That is, the state space is constructed in such a way that the correlation between a state and a given instruction can be maximum. Thus, through the task accomplishment process, the robot emerges a symbol (physical meaning) grounded by its actions. The method is applied to an indoor navigation and a soccer shooting tasks.

In the existing approaches, there have been some methods to construct the visual state spaces through

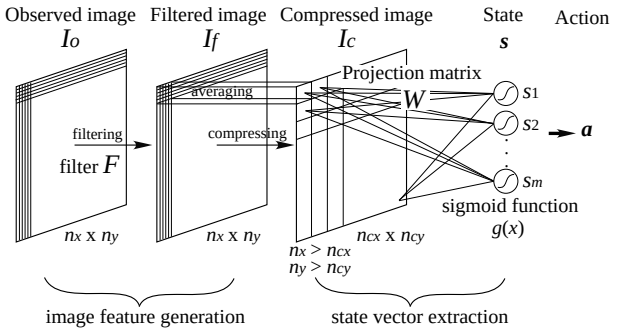


Figure 1: Image feature generation and action selection model

the task executions (e.g.(Nakamura 1998; Ishiguro, Kamiharako, & Ishida 1999)). These methods can construct the task-oriented state vector, but they don't focus on the image features. Our proposed method constructs the task-oriented visual state space and image feature which is useful for the selective attention.

The rest of the paper is organized as follows. First, the basic idea for image feature generation is described along with the learning formulation. Here, the projection matrix from the extracted image feature to the state vector is introduced to consequently determine the optimal action. Next, the experimental results are given to show the validity of the proposed method. Finally, discussion on the attention mechanism suggested from the current results is given towards the next step.

Image feature generation

A basic idea

Fig.1 shows the proposed model of the system for image feature generation and action selection. The reason why adopting two stage learning is that we expect the former more general and less task specific while the latter vice versa. In other words, at the image feature generation stage, the interactions between raw data are limited inside local areas while the connections between the image features and the states spread over the entire space to represent more global interactions. A similar structure can be found in the synapse connections in our brain, where the retinal signals geometrically close to each other are mapped to nearby regions in the early visual processing area while the post-processing and therefore more abstracted information is spread out the whole brain via a number of synapse connections (ex. (Fellman & Essen 1991)).

We prepare the image filter F to generate the image features. The robot estimates its state $\boldsymbol{s}$ from the filtered image I_f and decides the action appropriate to the current state $\boldsymbol{s}$. In order to avoid *curse of dimension*, we compress the filtered image I_f into I_c from which the state vector is extracted by a projection matrix W. We can regard W as a kind of attention mechanism because it connects the filtered image I_c to the state space, that is, it tells which part in the view is more important to estimate each state, and finally to decide the optimal action. Therefore, the problem is how to learn F and W.

In order to reflect the task constraints, we use the supervised successful instances (a training set). F and W are computed by minimizing the conditional entropy of the action given the state on the training set.

In this paper we prepare a 3×3 spatial filter F_s and a color filter F_c as follows:

- a 3×3 spatial filter $F_s \in \Re^{3\times 3}$:

$$\begin{aligned}\bar{I}_{xy} = \; & f_{s11}I_{x-1y-1} + f_{s12}I_{xy-1} + f_{s13}I_{x+1y-1} \\ & + f_{s21}I_{x-1y} \quad + f_{s22}I_{xy} \quad + f_{s23}I_{x+1y} \\ & + f_{s31}I_{x-1y+1} + f_{s32}I_{xy+1} + f_{s33}I_{x+1y+1},\end{aligned}$$

$$I_{fxy} = g\left(\bar{I}_{xy}\right).$$

- a color filter $F_c \in \Re^3$:

$$\bar{I}_{xy} = f_{c1}I_{rxy} + f_{c2}I_{gxy} + f_{c3}I_{bxy},$$

$$I_{fxy} = g\left(\bar{I}_{xy}\right),$$

where x and y denote the position of the pixel, I, I_r, I_g and I_b the gray, red, green and blue components of the observed image, respectively, and $g(\cdot)$ a sigmoid function. For example, the following F_s and F_c represent a vertical edge filter and a brightness one, respectively.

$$F_s = \begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix},$$

$$F_c = \begin{pmatrix} 0.2990 & 0.5870 & 0.1140 \end{pmatrix}^T.$$

Learning method

In the teaching stage the robot collect the i-th pair

$$T_i = < I_{oi}, \boldsymbol{a}_i >,$$

where I_o is the observed image, $\boldsymbol{a} \in \Re^l$ is the supervised robot action executed after the robot observes I_o and i denotes the data number. In the case of a mobile robot, l is two.

The state of the robot $\boldsymbol{s} \in \Re^m$ is extracted by $W \in \Re^{m\times n_{cx}n_{cy}}$. Let $\boldsymbol{i}_c \in \Re^{n_{cx}n_{cy}}$ be the one dimensional representation of I_c, then

$$\boldsymbol{s} = \boldsymbol{g}\left(W\boldsymbol{i}_c\right),$$

where $\boldsymbol{g}(\cdot)$ is a vector function of which components are sigmoid functions.

To evaluate F and W, we use the conditional entropy $H(\boldsymbol{a}|\boldsymbol{s})$:

$$H(\boldsymbol{a}|\boldsymbol{s}) = -\int p(\boldsymbol{s}) \int p(\boldsymbol{a}|\boldsymbol{s}) \log p(\boldsymbol{a}|\boldsymbol{s}) d\boldsymbol{a} d\boldsymbol{s},$$

where $p(\cdot)$ denotes the probabilistic density. To approximate $H(\boldsymbol{a}|\boldsymbol{s})$, we use the risk function R (Vlassis, Bunschoten, & Kröse 2001).

$$\begin{aligned} R &= -\frac{1}{N}\sum_d^N \log p(\boldsymbol{a}_d|\boldsymbol{s}_d) \\ &= -\frac{1}{N}\sum_d^N \log \frac{p(\boldsymbol{a}_d, \boldsymbol{s}_d)}{p(\boldsymbol{s}_d)}, \end{aligned}$$

where N is the size of the teaching data set. To model $p(\boldsymbol{a}, \boldsymbol{s})$ and $p(\boldsymbol{s})$, we use the kernel smoothing (Wand & Jones 1995).

$$p(\boldsymbol{s}) = \frac{1}{N}\sum_{q}^{N} K_s(\boldsymbol{s}, \boldsymbol{s}_q),$$

$$p(\boldsymbol{a}, \boldsymbol{s}) = \frac{1}{N}\sum_{q}^{N} K_a(\boldsymbol{a}, \boldsymbol{a}_q) K_s(\boldsymbol{s}, \boldsymbol{s}_q),$$

where

$$K_s(\boldsymbol{s}, \boldsymbol{s}_q) = \frac{1}{(2\pi)^{m/2} h_s^m} \exp\left(-\frac{||\boldsymbol{s} - \boldsymbol{s}_q||^2}{2h_s^2}\right),$$

$$K_a(\boldsymbol{a}, \boldsymbol{a}_q) = \frac{1}{(2\pi)^{l/2} h_a^l} \exp\left(-\frac{||\boldsymbol{a} - \boldsymbol{a}_q||^2}{2h_a^2}\right),$$

h_s and h_a are the width of the kernels. R can be regarded as the Kullback-Leibler distance between $p(\boldsymbol{a}|\boldsymbol{s}_d)$ and a unimodal density sharply peaked at $\boldsymbol{a} = \boldsymbol{a}_d$. By minimizing R, we can bring $p(\boldsymbol{a}|\boldsymbol{s})$ close to the unimodal density, that is, the robot can uniquely decide the action $\boldsymbol{a}$ from the state $\boldsymbol{s}$.

Using the steepest gradient method, we obtain a pair of F and W which minimize R:

$$F \leftarrow F - \alpha_f \frac{\partial R}{\partial F}, \quad W \leftarrow W - \alpha_w \frac{\partial R}{\partial W},$$

where α_f and α_w are the step size parameters.

After learning the robot executes the action $\boldsymbol{a}$ derived from its state $\boldsymbol{s}$ computed from the observed image as follows:

$$\boldsymbol{a} = \arg\max_{\boldsymbol{a}'} p(\boldsymbol{a}'|\boldsymbol{s}).$$

To find the maximum value, we adopt a coarse-to-fine search strategy.

Experiments

Task and assumptions

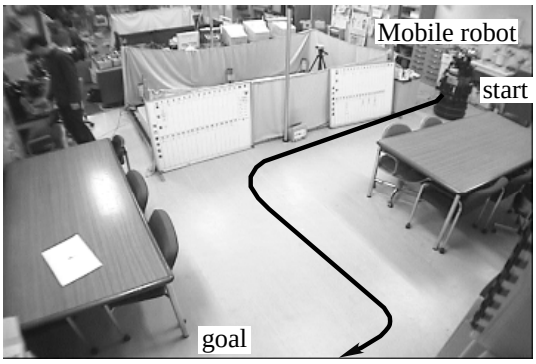
We applied the proposed method to an indoor navigation task with the Nomad mobile robot (Fig.2(a)) and a shooting ball task of the soccer robot (Fig.2(b)). The mobile robot shown in Fig.2(a) is equipped with stereo cameras and we use only the left camera image. The soccer robot shown in Fig.2(b) is equipped with a single camera directed ahead. The size of observed image is 64×54 and the values of I, I_r, I_g and I_b are normalized to [0 1]. The robots can execute translational speed v and steering speed ω independently, so the action vector is represented as

$$\boldsymbol{a} = (v, \omega)^T,$$

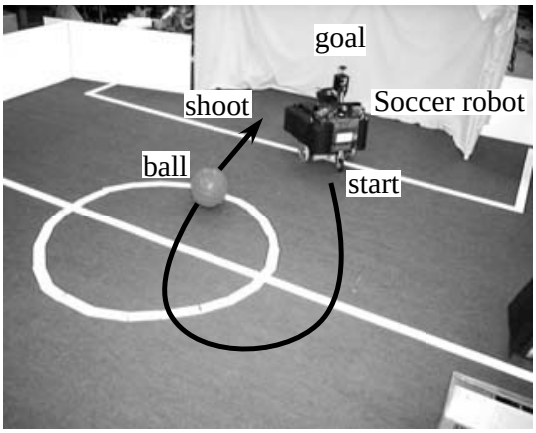
where v and ω are normalized to [-1 1], respectively. We define the size of the compressed filtered image as 8×6 and the dimension of state as $m = 2$. The sigmoid function g is

$$g(x) = \frac{1}{1 + \exp\left(-\frac{x-\theta}{c}\right)},$$

where $\theta = 0.0$ and $c = 0.2$.



(a) Task 1



(b) Task 2

Figure 2: Task

Learning results

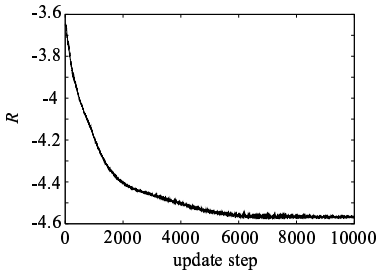
At the teaching stage, we gave 158 pairs of images and actions in the task 1, and 100 pairs in the task 2. In each task we tested the two models (Fig.1) with a spatial filter F_s and a color filter F_c, separately. We initialized the components of W by random small number and

$$F_s = \begin{pmatrix} 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 \end{pmatrix} \quad \text{(smoothing)},$$

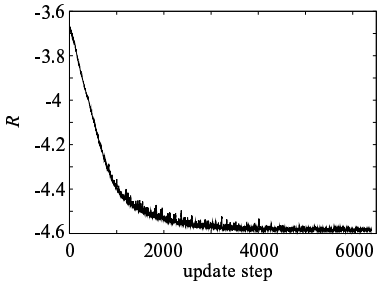
$$F_c = (\ 0.2990 \quad 0.5870 \quad 0.1140\ )^T \quad \text{(brightness)}.$$

Task 1: simple navigation Fig.3 shows the changes of R in the case of F_s and F_c models. F and W are learned so as to decrease R. Fig.4 shows the distributions of the state on the teaching data set in the case of the model with F_s. To show the relation between the states and actions, we labeled the action indices as follows:

- $v \geq 0.6$: forward,



(a) Model with F_s



(b) Model with F_c

Figure 3: Learning curves of R

- $v \leq -0.6$: backward,
- $-0.6 < v < 0.6$ and $\omega < 0.0$: right turn, and
- $-0.6 < v < 0.6$ and $\omega > 0.0$: left turn.

As we can see from these figures, the state space can be roughly classified in terms of actions. That is, the state space is constructed so that the correlation between an action and a class of states can be maximized. However it seems difficult to reveal a physical meaning from this relationship.

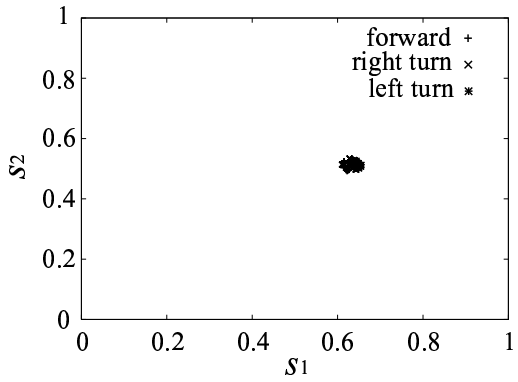
The generated F_s and F_c are shown below:

$$F_s = \begin{pmatrix} -0.8915 & -0.5995 & -0.06528 \\ -0.9696 & -0.4790 & 1.357 \\ -0.2482 & 0.1021 & 2.756 \end{pmatrix},$$

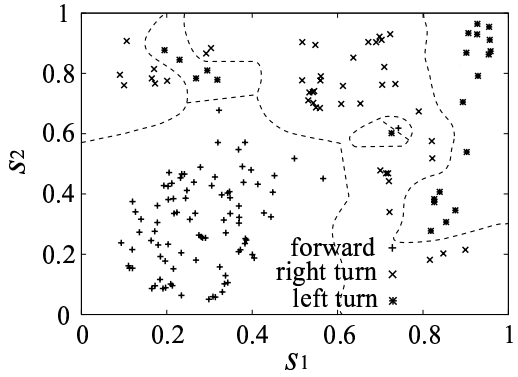
$$F_c = \begin{pmatrix} -0.4233 & 1.464 & -0.1718 \end{pmatrix}^T.$$

Figs.5 and 6 show the examples of the filtered images. As we can see from Fig.5, F_s shows the characteristic to extract vertical and horizontal edges. However F_c does not show remarkable characteristic because there are not salient color objects in the environment (our laboratory). Intuitively, it implies that the generated F_s is good at a navigation task of a mobile robot.

Task 2: shooting a ball The generated F_s and F_c are shown below. The examples of the filtered image



(a) Initial state space



(b) Learned state space

Figure 4: State distributions

are shown in Figs.7 and 8.

$$F_s = \begin{pmatrix} -3.384 & -1.953 & -1.686 \\ 0.3491 & -1.350 & 0.5363 \\ 1.656 & -1.208 & 5.223 \end{pmatrix},$$

$$F_c = \begin{pmatrix} 1.836 & 1.616 & -4.569 \end{pmatrix}^T.$$

F_s shows the characteristic to extract horizontal edges (see Fig.7). F_c emphasizes the red ball and yellow goal but inhibits the white line and wall. This is equivalent to a characteristic of a reversed U component of YUV image. The generated F_c is good at a soccer robot task in the colored soccer field.

Learned behavior

To verify the validity of the learned model, we applied the model with F_s (task 1) to a navigation task of the Nomad mobile robot (see Fig.2(a)). Fig.9 shows a sequence of the acquired behavior. The estimated states

(a) Io (b) If (c) Ic

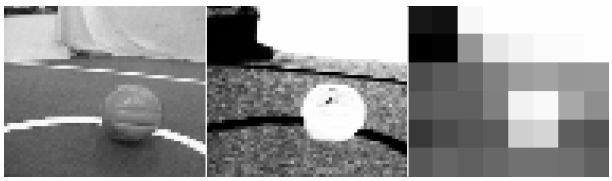
Figure 5: Example of the filtered image with F_s

(a) Io (b) If (c) Ic

Figure 6: Example of the filtered image with F_c

(a) Io (b) If (c) Ic

Figure 7: Example of the filtered image with F_s



(a) Io (b) If (c) Ic

Figure 8: Example of the filtered image with F_c

in this experiment are not exactly coincident with the states computed from the teaching data set, but the robot accomplished the task. Hence it implies that this model is an effective representation for the task and environment.

Discussion and future work

In this paper we proposed the method to generate an image feature and to learn a projection matrix from the image feature to the state, that suggests which part in the view is important, that is, a gaze selection by visio-motor mapping. The generated image features are appropriate for the task and environment. Also the acquired projection matrices give appropriate gaze selection for the task and environment. To show this, we illustrate the absolute value of W acquired in the model with F_c in the task 2. Figs.10(a) and (b) show the values of components of W related to the first and second components of the state $\boldsymbol{s}$, respectively. In these figures brighter pixels are more closely related to the state vector, that is, the robot gazes this parts in the view. Therefore we can regard that a projection matrix gives a gaze selection.

In our experiments, the instructor gave a motor command at every situation although he or she liked to give more abstracted instructions such as "avoid an obstacle" or "make a detour" since the robot dose not have any mechanism to interpret such instructions. Fig.4 (a) indicates this situation where there is no correlation between action and state. However, through the proposed learning process, the robot constructed the state space as shown in Fig.4 (b) where each action almost corresponds a cluster of states. This may imply that the physical meaning of instructions (the symbol) intended by the instructor can be interpreted (grounded) by the robot.

In this paper we defined the dimension of the state vector as two heuristically. However, as a result, this number was appropriate. Fig.11(a) shows the relationship between the number of dimension and R in the case of learning the model with F_s in the task 1, and Fig.11(b) shows the relationship between the number of dimension and the action estimation error. This error shows the sum of error norms between the estimated action and the supervised action. From these figures we can see that the necessary number of dimension is two to estimate the action from the state.

In the sequence of Fig.9 there are some cases that the robot decides the actions with relatively low probability $p(\boldsymbol{a}|\boldsymbol{s})$, that is, the robot is not so sure about its action decision. Therefore, it seems necessary for the robot to select multiple image features from the image feature set to accomplish more complicated tasks. Now, we are investigating how to integrate the proposed method and the image feature selection method based on the information theoretic criterion (Minato & Asada 2000).

References

Arbel, T., and Ferrie, F. P. 1999. Viewpoint selection by navigation through entropy maps. In *Proceedings of the Seventh International Conference on Computer Vision*, 248–254.

Fellman, D. J., and Essen, D. C. V. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex* 1:1–47.

Ishiguro, H.; Kamiharako, M.; and Ishida, T. 1999. State space construction by attention control. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1131–1137.

Kröse, B. J. A., and Bunschoten, R. 1999. Probabilistic localization by appearance models and active

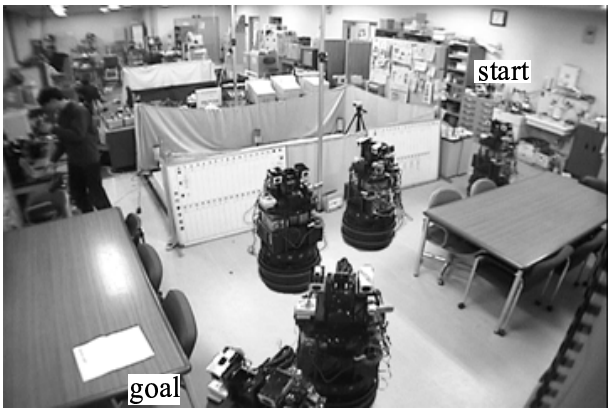
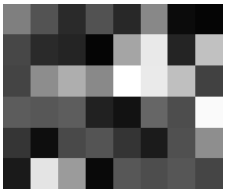
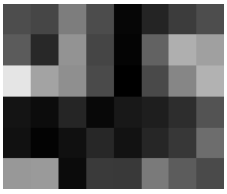


Figure 9: Acquired behavior

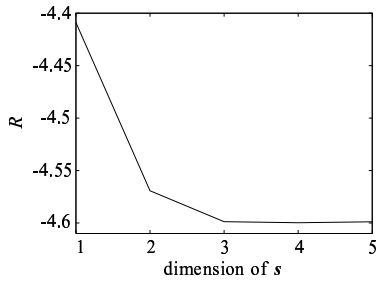


(a) The first component of $\boldsymbol{s}$

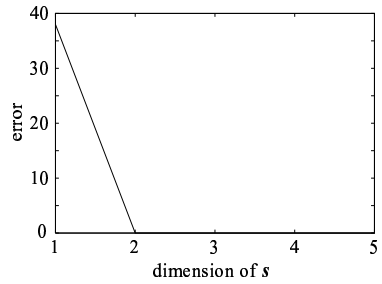


(b) The second component of $\boldsymbol{s}$

Figure 10: Projection matrix W



(a) R



(b) Estimation error

Figure 11: Effect of the dimension of the state vector

vision. In *Proceedings of the 1999 IEEE International Conference on Robotics and Automation*, 2255–2260.

Laar, P., and Gielen, S. 1997. Task-dependent learning of attention. *Neural Networks* 10(6):981–992.

Minato, T., and Asada, M. 2000. Selective attention mechanism for mobile robot based on information theory. In *Proceedings of the 18th Annual Conference of the Robot Society of Japan*, 811–812. (in Japanese).

Nakamura, T. 1998. Development of self-learning vision-based mobile robots for acquiring soccer robots behaviors. In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, 2592–2598.

Nayar, S. K.; Murase, H.; and Nene, S. A. 1996. *Parametric Appearance Representation in Early Visual Learning*. Oxford University Press. chapter 6.

Takeuchi, Y.; Ohnishi, N.; and Sugie, N. 1998. Active vision system based on information theory. *Systems and Computers in Japan* 29(11):31–39.

Thrun, S. 1998. Bayesian landmark learning for mobile robot localization. *Machine Learning* 31(1).

Treisman, A., and Gelade, A. 1980. A feature integration theory of attention. *Cognitive Psychology* 12:97–136.

Vlassis, N.; Bunschoten, R.; and Kröse, B. 2001. Learning task-relevant features from robot data. In *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, 499–504.

Wand, M. P., and Jones, M. C. 1995. *Kernel Smoothing*. Chapman & Hall.

Wolfe, J. M.; Cave, K. R.; and Franzel, S. L. 1989. Guided search: An alternative to the feature integration model. *Journal of Experimental Psychology: Human Perception and Performance* 15(3):419–433.