Detection and Categorization of Facial Image through the Interaction with Caregiver

Masaki Ogino*, Ayako Watanabe† and Minoru Asada*
*JST ERATO Asada Synergistic Intelligence Project
†Graduate School of Engineering, Osaka University,
2-1 Yamadaoka, Suita, Osaka, Japan
Email: [ayako.watanabe, asada]@ams.eng.osaka-u.ac.jp, [ogino, asada]@jeap.org

***Abstract*—This paper models the process of Applied Behavior Analysis (ABA) therapy of autistic children for eye contact as the learning of the categorization and preference through the interaction with a caregiver. The proposed model consists of the learning module and visual attention module. The learning module learns the visual features of higher order local autocorrelation (HLAC) that are important to discriminate the visual image before and after the reward is given. The visual attention module determines the attention point by a bottom-up process based on saliency map and a top-down process based on the learned visual feature. The experiment with a virtual robot shows that the robot successfully learns visual features corresponding to the face firstly and the eyes afterwards through the interaction with a caregiver. After the learning, the robot can attend to the caregiver's face and eyes as autistic children do in the actual ABA therapy.**

***Index Terms*—categorization, autism, ABA therapy, eye contact, HLAC, weakly supervised learning**

I. INTRODUCTION

In adaptive communication, it is necessary to learn what information an agent should pay attention to. Especially in visual information, this includes two issues: (1) to find appropriate visual pattern and (2) to give the meaning in communication to the acquired pattern. However, these two kinds of issues are difficult to separate clearly, and it is necessary to design a method to find important visual information appropriately depending on the situations.

This is closely related to cognitive development of human. Infants are known to have the preference to the face-like pattern just after their born, and this ability is believed to be innate [1] [2]. This ability matures as they grow up so that they can learn to distinguish and categorize different people in terms of various points [3]. However, recently, Fasel et al. mentioned that only 6 minutes are enough to collect the data to train the detector for face-like image and it is still questioned that the facial preference is innate [4]. On the other hand, some people with autism spectrum disorders do not show the preference to attend human faces. This is thought to be one of the causes why they fail to learn how to communicate with others. Actually, some therapy emphasizes in training autism children to look at the other's face and they admit the improvement in social cognitive ability to some extent [5]. In these two examples, categorization through interaction is an interesting problem for modeling as the constructivism approach [6].

The problem of categorization through interaction can be modeled as the categorization of data in one sensor based on another sensor in multi-modal sensory inputs. In this paper, the therapy of children with autism spectrum disorders for acquiring eye contact based on ABA (Applied Behavior Analysis) is modeled as such an example of the categorization learning through interaction. With this model, the communication interaction promotes the categorization of visual information. Then, the new categorization affects the preference of the agent behavior in interaction, which promotes the more categorization of visual information afterwards.

This paper is organized as follows. First, the procedure that are actually carried out in the therapy for children with autism spectrum disorders is introduced. Second, the learning model to explain the learning process of children with autism spectrum disorders is proposed. Third, experimental results with virtual robot are shown. Finally, discussion and conclusions are given.

II. CATEGORIZATION THROUGH COMMUNICATION

It is observed that some people with autism spectrum disorders have a weak preference to the eyes of others [7]. This distinctive difference of the preference of the attention from usual people is suspected as one of the causes why children with autism spectrum disorders fail to acquire the communication skills. However, some children with autism spectrum disorders can acquire the preference to the eyes through the Applied Behavior Analysis (ABA) therapy. ABA therapy was firstly developed by Lovaas, and it is reported that social skills of children with autism spectrum disorders are improved to some extent through ABA therapy [8]. Although many techniques are proposed in ABA therapy, the basic idea is to classify the social behaviors into the behavior elements and reinforce each behavior element by the reward. Fig. 1 shows the process of the ABA therapy to reinforce the eye contact behavior to children with autism spectrum disorders that are actually applied by the therapist [5]. In the first stage, an autistic child attends only to the object that he/she wants, such as a favorite toy, and does not attend to the therapist. In the training phase, the therapist does not give the object even if the child reach his/her hand to the object until the child happens to move their gaze to the face of the therapist.

As the therapy proceeds, the autistic child learn to look at the face and afterwards the eyes of the therapist when he/she feels demands. In the actual therapy, this training is regarded as one of the most important stages for the following therapy for various social cognitive skills [5].

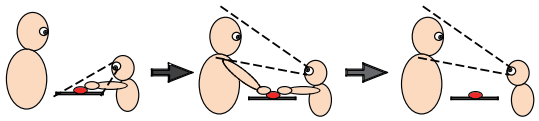

Fig. 1. The process of the ABA therapy for eye contact

In this process, it is necessary for the autistic child to realize that the visual pattern of the face and the eyes have some important information and for the reward and to categorize these patterns as such, even though there are many possible visual features in his/her visual field. The categorization of visual features are thought to affect to the visual attention and vise verse in the communication interaction. Thus, in this paper, we propose a learning model that categorizes the important visual feature based on the reward information to analyze the relationship between the visual attention and the categorization in the communication.

III. LEARNING KEY FEATURES BY REWARD

A. Overview

Fig. 2 shows an overview of the proposed system. The system consists of the learning system and the vision system. The learning system records the images and separates them into two groups depending on the timing before and after the reward is given. Then the visual features are learned so that they discriminate the images into two groups well. In the vision system, input images are processed with the top-down and bottom-up processes to calculate the candidates of the attention points in the camera image that the agent attends to. The bottom-up process selects the attention points by the saliency level of the image, and the top-down process selects the attention points by the similarity to the learned visual features. In the following, the details of the system are explained.

B. Segmental HLAC Features

1) HLAC: Higher order local autocorrelation (HLAC) feature is proposed by Otsu and Kurita [9]. The N-th order autocorrelation functions with displacements $(a_1, a_2, ..., a_n)$ from the reference point r are calculated as

$$x_N(a_1, a_2, \cdots, a_N) = \int I(\mathrm{r})I(\mathrm{r}+a_1)\cdots I(\mathrm{r}+a_N)dr, \quad (1)$$

where $I(\mathrm{r})$ is the intensity at the position r. In HLAC, the number of these autocorrelation functions, N, is usually limited up to the second ($N = 0, 1, 2$), and the range of the displacements is limited to the local 3 x 3 window as shown in Fig. 1. The element of HLAC features (corresponding to each local pattern) is calculated as the integration of the autocorrelation

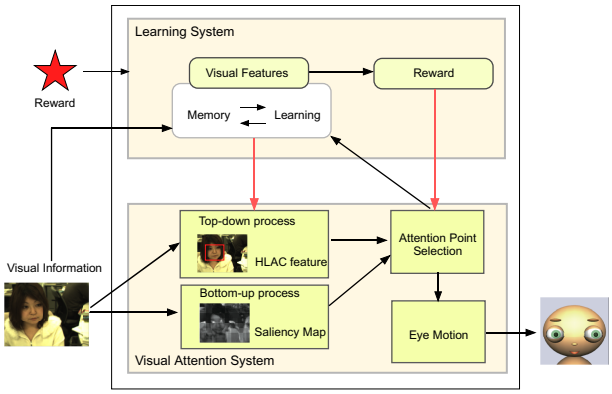

Fig. 2. An overview of the proposed system

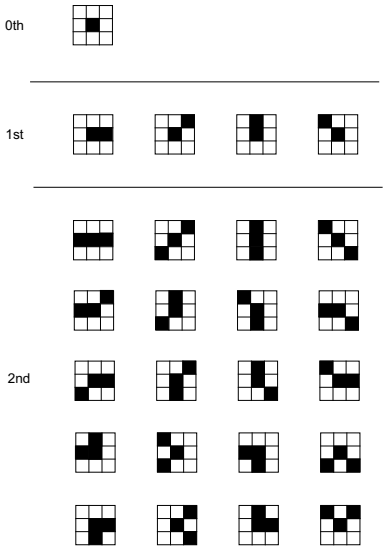

Fig. 3. Local mask patterns for computing HLAC features

in each pixel all over the image. HLAC features have shift-invariant because autocorrelation function is shift-invariant. As HLAC keeps the generality in the image, it is also used for facial expressions by weighting HLAC vectors [10].

C. Evaluation of HLAC features for facial detection

In the existing studies using HLAC features for recognition, HLAC features are usually weighted by linear discriminant analysis [11] [12] or Fisher weight maps [10] for improving the recognition rates. However, in weakly supervised learning, it is difficult to acquire the objective data. Only the labeled (knowing the object is included but not knowing where it is) and unlabeled images are available. In this problem setting, it is difficult to calculate the appropriate weights in advance. Before using the HLAC features for weakly supervised learning, we tested whether raw HLAC features have the appropriate characteristics for detecting the face.

Before calculating the HLAC features, the captured camera image is processed by the Canny filter to extract the edge features.

Whether the segment x includes an object (in this case, the face) or not is evaluated by the distance to the reference HLAC feature,

$$d = \|\boldsymbol{h}^x - \boldsymbol{h}^{ref}\| \tag{2}$$

and the segment is labeled depending on the distance,

$$l_x = \begin{cases} 1 & if \quad d < \phi \\ 0 & else \end{cases} \tag{3}$$

In the first test, the robustness for detection is examined in the experimental environment. Fig. 4 shows the uniqueness of the HLAC feature corresponding to the face when the segment including the face is given. The region colored in red is the area that is evaluated as including the front face. The image size is 640×480 and the segment size (surrounded by white lines) is 200×200. The threshold is set as $\phi = 40$ (this value is comparable to the learning result explained in the next section).

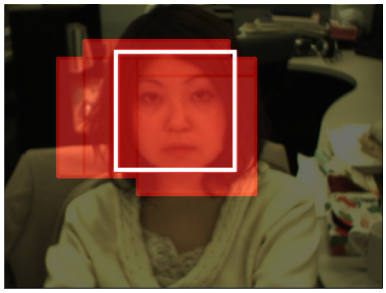

Fig. 4. Face detection in complex environment

In the second test, the effects of the face orientation on the HLAC features are examined. Fig. 5 shows the distance of the HLAC features between the faces with various orientations and the front face (4). In this figure, 8 is the average and the standard deviation of the distance to non-face segments selected randomly in the experimental environment. The image mentioned in 8 is one of the examples in the environment except human face. This graph shows the possibility that we can set the threshold that separates the face including various orientations from the other environment and can set the threshold that separates the front face from the directed faces.

D. Finding the key feature

This section proposes a method that finds the key feature h_{ref} and the appropriate threshold ϕ to evaluate whether one segment includes the object or not, based on the data set: labeled images and unlabeled images. The proposed system is shown in Fig. 6. Note that in labeled image we do not know which segment includes the object (in this case, the face). The basic idea is to extract the segment including the object from the labeled images.

The main procedure is following.

1) Select one reference image $X^{ref} \in \mathbf{A}$ among the labeled images $\mathbf{A} = \{X_1^a, X_2^a, \cdots, X_N^a\}$.

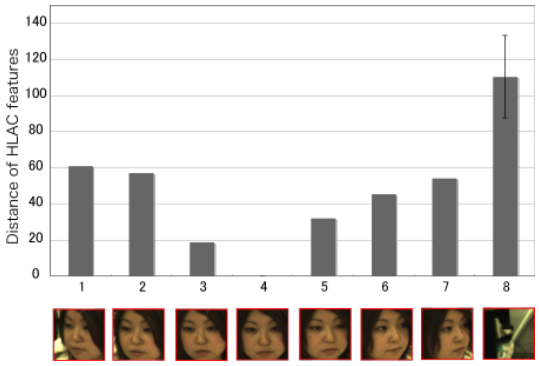

Fig. 5. The distances of HLAC features from the frontal to the oriented faces and the environment

2) Segment the reference image X^{ref} into the L image segments Z_i^{ref} $(i = 1...L)$.
3) For each reference image segment Z_i^{ref} and the possible threshold value ϕ, proceed the following procedures.
   a) Segment the labeled image $X_n^a \in \mathbf{A}$ and unlabeled images $X_n^b \in \mathbf{B}$ into L image segments, $Z_{n,\,i'}^a$, $Z_{n,\,i'}^b$.
   b) Calculate the distances, $d_{i\,i'}^a$, $d_{i\,i'}^b$, between each segment, $Z_{n,\,i'}^a$, $Z_{n,\,i'}^b$, and the reference segment, Z_i^{ref}.

$$d_{i\,i'}^{(a,b),n} = \|\,\mathbf{h}_{i'}^{(a,b),n} - \mathbf{h}_i^{ref}\| \tag{4}$$

   where $\mathbf{h}_{i'}^{(a,b),n}$ is the HLAC feature of the image segment $Z_{n,\,i'}^{a,b}$, and $\mathbf{h}_i^{ref}$ is the HLAC feature of the image segment Z_i^{ref}.
   c) Among the image segments in labeled and unlabeled images, select the segments, $Z_{n,\,i'_{min}}^a$, $Z_{n,\,i'_{min}}^b$, that has the minimum distance from the reference segment Z_i^{ref}.

$$\begin{aligned} i'^{a,b}_{min} &= \underset{i'}{argmin} \;\; \|\,\mathbf{h}_{i'}^{(a,b),n} - \mathbf{h}_i^{ref}\| \\ &= \underset{i'}{argmin} \;\; d_{i,\,i'}^{(a,b),n} \end{aligned} \tag{5}$$

   d) Calculate the distances between the image segments, $Z_{n,\,i'_{min}}^a$, $Z_{n,\,i'_{min}}^b$ and the reference image segment, $d_{i\,i'_{min}}^{a,\,n}$, $d_{i\,i'_{min}}^{b,\,n}$ as follows,

$$d_{i\,i'_{min}}^{(a,b),n} = \min_{i' \in X_n^a, X_n^b} d_{i\,i'}^{(a,b),n} \tag{6}$$

   e) Assign the label to the segment that mentions whether the segment includes the object or not,

$$l_{\,n}^{\,a,b} = \begin{cases} 1 & if \quad d_{i\,i'_{min}}^{(a,b),n} < \phi \\ 0 & else \end{cases} \tag{7}$$

   f) Evaluate the recognition rate by comparing the assigned label with the group label,

$$I_n^{a,b} = \begin{cases} 1 & if \quad l_{\,n}^{\,a} = 1 \quad or \quad l_{\,n}^{\,b} = 0 \\ 0 & else \end{cases} \tag{8}$$

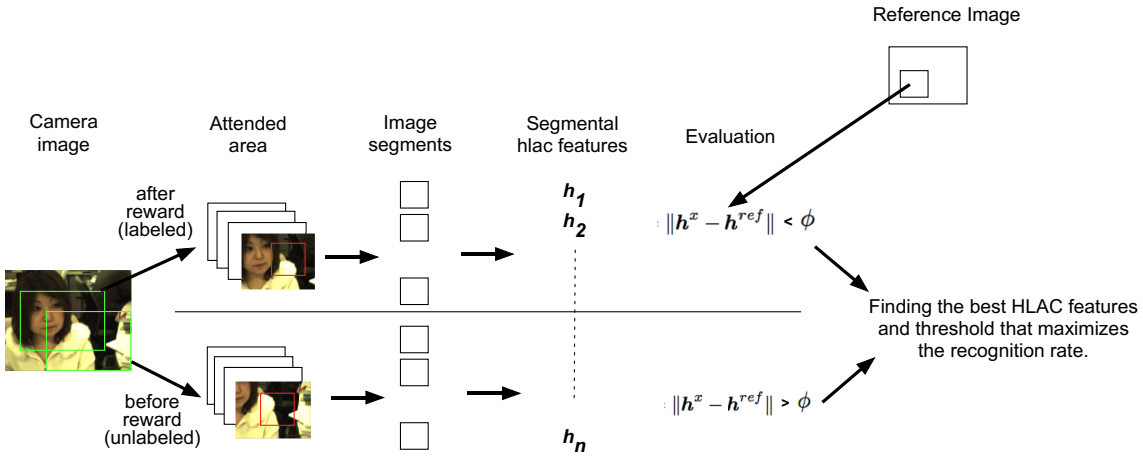

Fig. 6. Overview of the learning system for important visual features

$$\mu(\phi, h_i) = \frac{\sum_{n=1,a,b}^{N} I_n^{a,b}}{2N} \tag{9}$$

4) Find the set of HLAC feature and the threshold that makes the highest recognition rate.

$$(\phi_{max}, h_{i_{max}}) = \underset{i,\phi}{arg\max} \mu(\phi, h_i) \tag{10}$$

E. Visual attention system

The visual attention system decides the attention point from the candidates that are calculated in the bottom-up process (based on the saliency map [13]) and the top-down process (based on the learned HLAC features) (Fig. 7). The attention point calculated in the top-down process and in the bottom-up one is selected with the probability P and $(1-P)$, respectively. The probability P is updated depending on the reward as follows,

$$P_t = P_{t-1} + \alpha r_t \tag{11}$$

where P_t indicates the probability to select the candidate calculated in the top-down process when the t-th reward r_t is given. In the experiment, α is set to 0.2. Thus, the robot learns to attend to the points that are related to the reward. When the top-down process successes in learning the face or eyes pattern as reward-related information, the robot learns to show the preference to the face or the eyes. The attended area is defined as the square area whose center is the attended point. The size of the attended area is set to 480×320, two thirds of the size of camera image, 640×480. The images of the attended area are recorded as the labeled and unlabeled images.

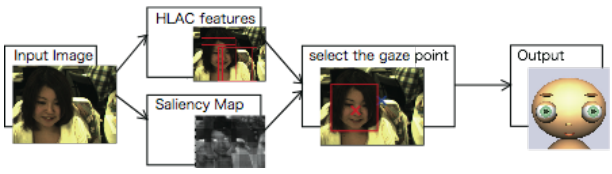

Fig. 7. Visual attention system

The robot has the demanding state. When the demanding state starts, the robot starts to record the attended areas as the unlabeled group until it receives the reward, and it records the attended areas after the reward is given as the labeled group (Fig. 8).

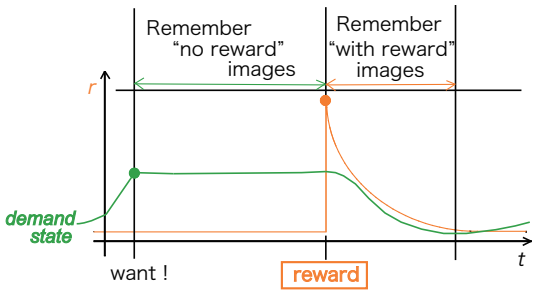

Fig. 8. The robot remembers images depending on the timing when it gets the reward.

F. Interaction procedure

During the learning, the size of the image segment for learning becomes smaller depending on the variance in the learned HLAC features, $v(\mathbf{h})$ (Fig. 9). In this paper, the sizes of the image segment is set to 300×300, 200×200 and 100×100 as the learning stage proceeds.

In the first interaction, the robot starts the learning without knowing the facial pattern. The attention point is only calculated by the bottom-up process using saliency map. The caregiver gives the reward to the robot when the robot happens to see the caregiver. After the several learning, the variance of the selected visual features become small and the robot begin to look at the face. Thus, the characteristic of the labeled and unlabeled groups at this stage becomes different from that at the first stage. When the robot acquires the HLAC feature of the facial pattern, the robot often attends the caregiver even if the caregiver does not see the robot. Therefore, it is expected that the main difference in the images between the unlabeled

and labeled group becomes the appearance of the face when the caregiver look and does not look at the robot.

Fig. 9. Stream of Learning

IV. EXPERIMENTAL RESULTS

A. Experimental setting

Fig. 10 shows the experimental setting. The human caregiver interacts with the virtual robot in the screen. The IEEE 1394 camera is set on the display. The size of the camera image is 640×480 pixels. The gaze direction of the virtual robot is calibrated with attention point in the camera image in advance. The reward is given to the robot by the caregiver with the keyboard when the caregiver feels that the virtual robot looks her.

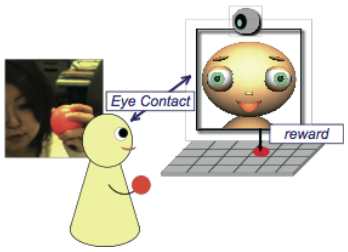

Fig. 10. The environment of the experiments

B. Learning through interaction

In the first experiment, it is shown that the proposed system can learn the visual features corresponding to the caregiver's face and eyes based on the reward information. Fig. 11 shows the learned image segments that are evaluated as important for discrimination of the labeled and unlabeled images. The green square indicates the attended area, and the red squares indicate the areas that have high recognition rate μ in the eq. (9). As mentioned in the figure, as the learning stage proceeds, the selected image patterns change from the face to the eyes.

Fig. 12 shows the attended points of the robot before and after the learning. Before the learning, the attention points move around corresponding to the slight changes of the caregiver's postures and the environment. After the learning, the robot attention points gathers around the eyes of the caregiver by virtue of the learned visual feature.

The variance of the attention points to the eye are calculated to evaluate how the attention points gathers around the eyes. The attention points are recorded during 1 minute and the distance from the attention points to them are calculated in the pixel unit. Fig. 13 shows the resultant variances before and after the learning. This graph shows the attention points converges to the eye pattern well in virtue of the learning.

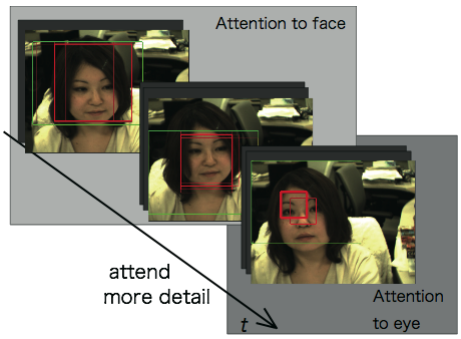

Fig. 11. Learned image segments in each learning stage

(a) Before learning (b) After learning

Fig. 12. The distribution of attention points of the robot before and after the learning

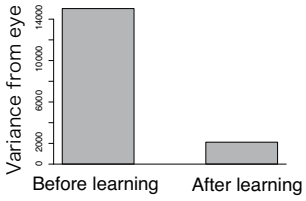

Fig. 13. The variance of the distance of the attention points from the eye before and after the learning

C. The effects of the motion on the learning

In the second experiment, the effects of the saliency on the learning of the preference is investigated. It is reported that the attention of some children with autism spectrum disorders is less affected by the motion information [14] and it is thought that this might be the cause of the failure of children with autism spectrum disorders to acquire the communication skills. We investigate how the learning of the visual feature and the preference of the attention are affected in case that the saliency of motion is not utilized.

Fig. 14 shows difference between the attention points based on the saliency map with and without motion saliency. With motion saliency, the face is salient to some extent even before the learning because of the eye blinks and the head motion, and the attention can be easily led to the face by the appealing motion like "Look at me!" that usual people often do to an

infant. On the other hand, without motion saliency, once the attention points fall to the edges or color in the environment, it is difficult to lead the attention to the human.

(a) With motion information (b) Without motion information

Fig. 14. Difference of Gaze direction using Saliency Map

This attention difference affected the learning time. Fig. 15 shows the learned visual features in each learning time. With the motion saliency, the robot can learns the facial pattern with only once. Without the motion saliency, the robot still can learn the facial pattern but it took 3 times learning time. These results show that the attention control with the motion saliency promotes the learning of the important visual features.

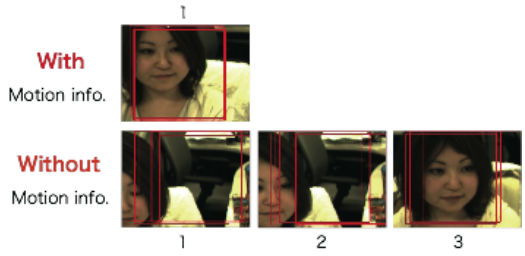

Fig. 15. Difference of learning time

V. Discussion and Conclusion

This paper models the process of the therapy for children with autism spectrum disorders based on ABA as the learning problem of the categorization and the preference through the interaction. The proposed learning system can acquire the visual features in the camera image according to the timing when the reward is given. The robot decides an attention point through the bottom-up process based on the saliency map and the top-down process based on the learned visual feature. In the experiment, the caregiver gives the reward to the robot when she feels the robot looks at her. First the robot acquires the visual feature of the caregiver's face comparing the images that include the caregiver's face with the images that do not include. Afterwards, the robot often looks at the caregiver with the virtue of the learned visual feature. Thus, in turn, it compares the images of oriented face of the caregiver with those of the front face, and can acquire the eye pattern. This is a good example of the interactions among the communication interaction, categorization and the preference.

This paper also reveals that it is important to change the attention based on the motion saliency for promoting the learning of important visual feature. It is said that children with autism spectrum disorders are little affected by the motion saliency. However, the mechanism of this lack of attention to the motion is not made clear yet. The likely causes are that they cannot detect the motion saliency as assumed in this paper or that they cannot move their attention to the detected salient point. However, considering that some people with autism spectrum disorders are well known to look at the regular motions with enthusiasm, the control does not seem to have a problem. Thus, it may be that they cannot release their attention and cannot back to the preparation state so that they can attend to the new salient points. The proposed system models the attention system very simply as the bottom-up process based on the saliency map and the top-down process based on the learned visual feature. In the future model, we are planning to make a more detailed attention model and to investigate the role of the attention in the communication learning so that the model can be compared with the observed phenomena in children with autism spectrum disorders .

References

[1] M. H. Johnson, S. Dziurawiec, H. Dllis, and J. Morton, "Newborns' preferential tracking of face-like stimuli and its subsequent decline," *Cognition*, vol. 40, pp. 1–19, 1991.

[2] M. H. Johnson, "Subcortical face processing," *Nature reviews*, vol. 6, pp. 766–774, October 2005.

[3] A. Slater and M. Lewis, Eds., *Introduction to Infant Development*. Oxford University Press, 2007.

[4] I. Fasel, N. Butko, and J. Movellan, "Modeling the embodiment of early social development and social interaction: Learning about human faces during the first six minutes of life," in *Society for Research in Child Development Biennial Meeting*, 2007.

[5] H. Eguchi, (personal communication) http://www.kids-power.net/.

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

[7] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Choen, "Defining and quantifying the social phenotype in autism," *Social Phenotype in autism*, vol. 159, pp. 895–908, 2002.

[8] O. I. Lovaas, "Behavioral treatment and normal educational and intellectual functioning in young antistic children," *Journal of Consulting and Clinical Psychology*, vol. 55, no. 1, pp. 3–9, 1987.

[9] N. Otsu and T. Kurita, "A new scheme for practical flexible and intelligent vision systems," in *Proceedings of IAPR Workshop on Computer Vision*, 1988, pp. 431–435.

[10] Y. Shinohara and N. Otsu, "Facial expression recognition using fisher weight maps," in *Proceedings IEEE 6th International Conference on Automatic Face and Gesture Recognition (AFGR 2004)*, 2004, pp. 499–504.

[11] T. Kurita, N. Otsu, and T. Sato, "A face recognition method using higher order local autocorrelation and multivariate analysis," in *Proceedings of the 11th IAPR International Conference on Pattern Recognition*, vol. 2, 1992, pp. 213–216.

[12] K. Hotta, T. Kurita, and T. Mishima, "Scale invariant face detection method using higher-order local autocorrelation features extracted from log-polar image," in *Proceedings in the Third IEEE International Conference on Automatic Face and Gesture Recognition*, 1998, pp. 70–75.

[13] L. Itti, C. Koch, and E. Niebur, "A model of sliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[14] F. Shic, B. Scassellati, D. Lin, and K. Chawarska, "Measuring context: The gaze patterns of children with autism evaluated from the bottom-up," *ICDL*, 2007.