Ugur, E., Nagai, Y., Oztop, E., and Asada, M. (Eds) Proceedings of Humanoids 2012 Workshop on Developmental Robotics: Can developmental robotics yield human-like cognitive abilities? November 29, 2012. Osaka, Japan

Using Depth to Increase Robot Visual Attention Accuracy during Tutoring

Christian I. Penaloza, Yasushi Mae, Kenichi Ohara and Tatsuo Arai
Graduate School of Engineering Science
Osaka University
1-3 Machikaneyamacho, Toyonaka, Japan
http://www-arailab.sys.es.osaka-u.ac.jp

***Abstract*—This paper explores the problem of attention models for robot tutoring as related to the cognitive development of infants. We discuss the factors that have an important influence in infants' attention and the way these factors can be taken into consideration to develop robot attention models that simulate infants' cognitive stimuli. In particular, we focus on the attention given to objects that appear closer to the infant when they are shown by an adult. Using the distance of an object as an important factor to increase visual attention, our model uses depth information along with the well-known Bottom-Up Visual Attention Model Based on Saliency (Itti & Koch, 2001) in order to increase attention accuracy even if non-salient feature objects are shown to the robot or if tutoring activity takes place under cluttered environments. Our model also considers the presence or absence of a human tutor to decide whether a tutoring activity might take place. Experimental results suggest that depth information is a key factor to emulate effective infants' attention.**

I. Introduction

One of the main objectives of researchers in the area of cognitive robotics is to design mechanisms to provide robots with human-like abilities in perception, decision making, reasoning, and action execution. Among the many challenges of developmental cognition for robots, attention is perhaps one of the most important challenges that needs to be addressed since it plays a very important role in the process of learning.

The study of attention of infants can provide important clues to develop systems that emulate this important ability. This is because even before babies with normal vision can talk or walk, they are able to perceive and parse their visual environment and are able to move their eyes and head to select visual targets (objects or people) [7]. Moreover, by observing the cognitive development of infants when they interact with their parents, it has been shown that infants' attention and learning are favorably influenced by factors such as motionese (e.g., exaggeration of parent's actions) [8] or contingent reactions (helping infants find proper association) [10]. In this direction, researchers have developed robotic systems that emulate attention and learning processes of infants by using socially guided exploration [2], dialogs [4], or motionese [6]. Human guided instruction plays a very important role in infants learning and can be simulated in robot systems by presenting a visual task or object within robot's visual field as shown in Fig. 1 (a).

Fig. 1. 1) Human Guided Robot Learning. 2) Controlled environment (salient object and plain background) commonly used for experiments.

One of the main difficulties of robot learning is the fact that robots do not know where to look at when observing a demonstration [6]. Researchers have proposed computer vision models of attention that enable the robot to selectively choose a relevant visual segment while ignoring others (e.g., [19]-[22]). The *Bottom-Up Visual Attention Model Based on Saliency* originally proposed by Itti & Koch [20] is perhaps one of the most used and widely accepted models of attention. This model proposes the idea that visual attention is attracted by salient stimuli that 'pop out' from their surroundings due to primitive features such as color, intensity and orientation. This model is commonly used in robotic systems to achieve visual attention during a tutoring activity (e.g., [6], [25]). However, robotic systems that use this model are usually evaluated with objects that have strong salient features or in controlled environments with plain backgrounds (e.g., Fig. 1 (b)). This certainly facilitates the tutoring activity but limits its applicability to experimental setups and cannot be used in real environments. Moreover, since the Bottom-Up Attention Model uses 2D images to find 'salient' features to define focus of attention, distance of the object is not considered due to the lack of depth (3D) information.

The distance of an object to the infant is also a very important factor that effects visual attention. Smith *et al.* [1] provide experimental evidence that demonstrates that visual attention is largely increased by bringing objects close to the child. Therefore, this factor should certainly be used to improve attention mechanisms for robots. In this paper, we put this concept into practice and developed an attention model that uses depth information along with the Bottom-Up Visual Attention Model Based on Saliency in order to

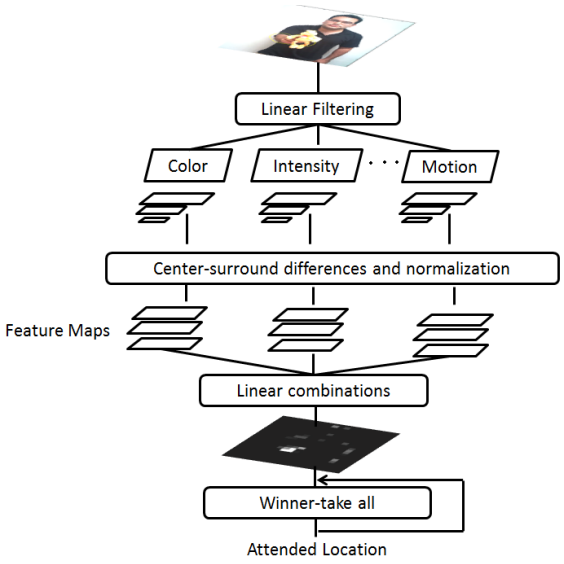

Fig. 2. Schematic of salience model proposed by Itti & Koch [20].

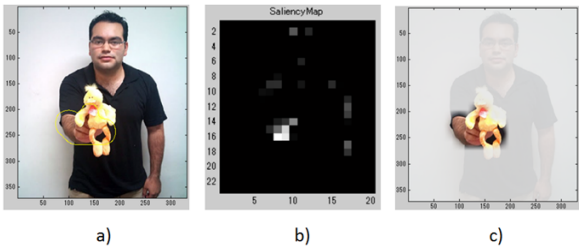

Fig. 3. Tutoring environment: a) Plain background - salient-feature object. b) Saliency Map c) Focus of attention. Attention model correctly locates object of interest.

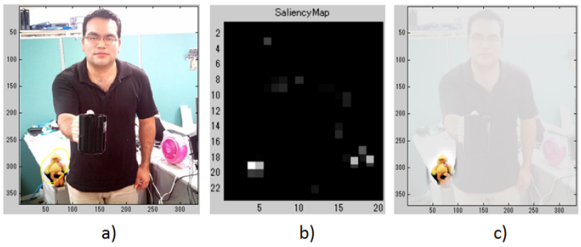

Fig. 4. Tutoring environment: a) Cluttered background, non-salient object, presence of distracters (salient-feature objects in background) b) Saliency Map c) Focus of attention. Attention model fails to locate the object of interest that is shown by the human tutor.

increase attention accuracy when the robot observes closer objects with or without salient features. We also take into consideration the presence or absence of a human teacher to activate the attention model when a tutoring activity takes place. Most importantly, our attention model can be used in real environments with cluttered backgrounds.

The remaining part of this work is organized as follows. Section II introduces a discussion of the Bottom-Up Visual Attention Model Based on Saliency and the need to use depth information for attention models. A description of our attention model is described in the subsequent section. Experiment design and results are given in section IV. Finally, in the last section we present some conclusive remarks.

II. A Bottom-Up Visual Attention Model Based On Saliency

Inspired by the behavioral and neuronal mechanism of primates, the *Bottom-Up Visual Attention Model Based on Saliency* uses the ”outstandingness” of primitive features of an image to be able to detect salient locations in a scene [20]. For example, a yellow object in a black background is detected as salient because of its distinctive color. A person moving to the right direction among other persons moving to the left direction is detected as salient with respect to motion direction.

This model is probably the most influential attention model, since it has been extensively used in many research fields including computer vision and robotics [11]. This model (Fig. 2) uses several concepts (e.g., feature map, saliency map) and proposes a well-structured process for calculation of the saliency map which defines attention focus. As a brief summary, multi-scale analysis of an input image is performed to evaluate five primitive features: color, intensity, orientation, flicker, and motion. Individual feature maps are combined to create a centralized saliency map that is used to identify the focus of attention. Refer to the original paper [20] for a more detailed description.

Contrary to the top-down attention model (an active scan of the visual field in search of a pre-specified object or stimuli), the bottom-up approach guides visual exploration focusing on the most salient stimuli - in a similar way babies do in early stages of development - and therefore it is more appropriate for emulating infant behavior. However, due to the native process of saliency computation from 2D images, the Bottom-Up Visual Attention Model Based on Saliency is unable to cope with attention focus based on depth information, which is also a key factor to effectively emulate infant attention.

A. Importance of Depth Information in Tutoring Activities

When an adult is tutoring an infant about an object or a particular task, the principle of *overt attention* (to place an object of interest at the center of visual field), along with the distance of the object are generally used to increase visual attention [13] - as demonstrated by experimental evidence of Smith *et al.* [1].

For tutoring activities, several researchers have emulated infants’ attention by successfully applying the Bottom-Up Visual Attention Model Based on Saliency in robotic systems. Since this model is meant to find ’salient’ features in the scene, most of the times the tutoring activity takes place in experimental environments (plain backgrounds scenes or use of objects with ’salient’ features such as bright colors), which certainly facilitate the learning task (i.e. Fig. 3). However, a real environment such as the one presented in Fig. 4 (a) (cluttered background, presence of multiple objects with salient features, or teaching an object that lacks salient features) presents a bigger challenge to the Bottom-Up visual model, which is unable to locate the object presented by the human tutor since other salient-featured objects are present in the

background. In this case, depth information plays a very important role in defining where the focus of attention should be located.

III. Attention Model using Depth

In order to deal with the difficulties presented in the previous section, our model uses depth information along with the Bottom-Up Visual Attention Model to be able to cope not only with feature saliency but also with object proximity. The main purpose is to emulate infants' attention when objects are presented - by a human tutor - at a close distance during a tutoring activity even if the objects lack strong feature saliency.

Attention models that use depth information to define focus of attention have been previously introduced by researchers (e.g., [14]-[17]) for scene analysis applications. In order to define saliency, mentioned techniques commonly use the 3D structural information of objects or object's relative position to other objects. Since none of these models takes into consideration the presence or absence of a person, it is difficult to implement them to emulate infant tutoring in robots because these models would encounter important difficulties such as focusing on the person vs object or detecting saliency effectively in extreme cluttered environments.

Since the main objective of this research is to emulate infants' stimuli, the particularity of our model is that we also take into consideration the presence or absence of a human teacher to activate the attention model, and we use *proxemics* theory according to Hall [23] to pre-define a distance range to which robots should pay attention when a tutoring activity takes place.

A. Development Process

During the tutoring activity, a human actively teaches an object to a robot by using the principle of *overt attention* (the object is presented within the visual field of the robot at a close distance). On the robot's behalf, our attention system that uses depth information along with the Bottom-Up Visual Attention Model is activated when a human teacher is found within its field of view. This is when the robot knows that a tutoring activity might take place. On the other hand, when a human teacher is not present or when there is no object close to the robot, only the Bottom-Up Visual Attention Model is activated.

Another way to look at our approach is by considering depth as an extra channel of the saliency map defined in the Bottom-Up model, but using a *binary* weight applied to the depth channel as described in Fig 5. This binary weight would be 0 when no human is present within the field of view of the robot and 1 otherwise - this can be considered as top-down influence of a human-detection. In other words, when a human is present, depth pixels within the personal space of the robot (if any) will be taken into consideration along with the pixels of salient features (color, intensity, orientation, flicker, and motion) of the Bottom-Up model. However, if no human is present, only the salient features of the Bottom-Up model are used to define attention location.

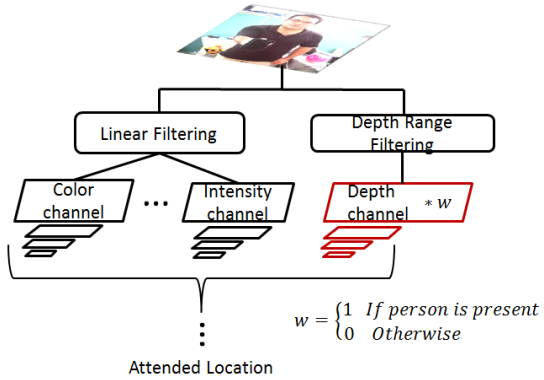

Fig. 5. Considering depth as an extra channel to saliency map defined in the Bottom-Up model, but using a *binary* weight applied to the depth channel.

Designation	*Specification*	*Usage*
Intimate distance	0 - 0.45m	Embracing or touching
Personal distance	**0.45 - 1.20m**	**Friends**
Social distance	1.20 - 3.60m	Acquaintances and strangers
Public distance	>3.60m	Public speaking

TABLE I
THE FOUR SPHERES OF PHYSICAL DISTANCE CORRESPONDING TO SOCIAL DISTANCE ACCORDING TO HALL [23].

In order to be able to recognize a human teacher, we used the built-in capabilities of our Kinect sensor through the Software Development Kit freely provided by Microsoft [24]. The main process of human body (pose) detection is explained in detail in [18]. As a brief overview of their method, the authors use a single depth image to accurately predict 3D positions of body joints by designing an intermediate body parts representation that maps the difficult pose estimation problem into a simpler per-pixel classification problem. Subsequently, they use a 'dictionary' of 3D pose proposals and find the closest match. Finally, they generate confidence-scored 3D proposals of several body joints by re-projecting the classification result and finding local modes.

In our model, once the human teacher is recognized, the depth-based attention is activated and the principle of depth-based saliency is performed. In this principle, a particular distance range is pre-defined and objects that appear within that range are given attention priority over objects that appear farther away from the robot even if they have stronger salient features than the ones of the object within the pre-defined distance range.

In order to define the most appropriate distance range for attention focus, we looked into social robotics literature and refer to the principle of *proxemics* — physical and psychological distancing from others. According to Hall [23], the four spheres of physical distance corresponding to social distance can be defined as described in Table I.

We chose personal distance as the most appropriate range for the tutoring activity, since objects within the intimate space appear too close to the camera and too far in the social space. The Kinect sensor has a depth distance limitation in which the minimum detection distance is 0.45m. Therefore, for our

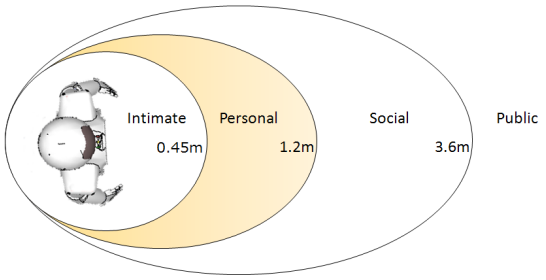

Fig. 6. Proxemics - From the four spheres of physical distance, personal distance was chosen as the most appropriate range for the tutoring activity.

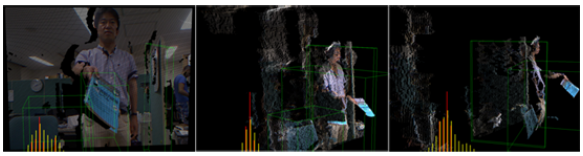

Fig. 7. Depth-RGB graphic representation as perceived by robot when human teacher demonstrates an object.

tutoring activity we defined the robot's personal space from 0.45m to 1.20m, as seen in Fig 6.

Kinect sensor provides valuable depth data that can be easily analyzed. Figure 7 shows a graphic representation of the visual Depth-RGB combination in which a human teacher is demonstrating an object to the robot. Our approach consists in extracting the RGB information corresponding to the object that appears within the personal distance range.

One way to detect objects within personal space is to perform depth-based thresholding. This involves estimating the depth value of each of the pixels that appear in the depth image and labeling those pixels whose z-value (depth) appears within the predefined distance range, as shown in Fig. 8 (d). Finally, target depth value pixels conform a pixel region that serves as a visual mask to the RGB input image to extract original color pixels of the attended object as observed in Fig. 8 (c).

The architecture of our attention model is described in Fig. 9. Our system integrates Bottom-Up Visual Attention Model with depth information by receiving input RGB and Depth images and use them to define whether a person is present or not, and decide the attention location based on the salient features of objects and object's distance to the robot.

IV. Experiment and Evaluation

A. Experimental Setting and Task

This section presents the experiment carried out to validate our attention model. The main objective of the experiment is to compare the attention accuracy during a tutoring activity in three cases: 1) using only the Bottom-Up Visual Attention Model, 2) using only depth-based attention model, and 3) using our model that combines both approaches.

The tutoring task consisted on a human volunteer presenting two types of objects to the robot: 1) objects with salient features (e.g. bright colors) and 2) objects with non-salient

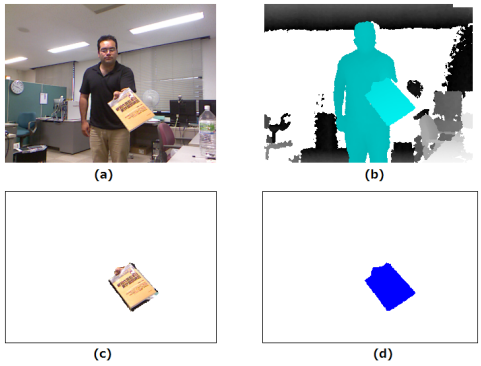

Fig. 8. Visual representation of our Depth-Based Attention Model: (a) Input RGB image, (b) Depth view - Human Detection, (c) Focus of Attention, (d) Personal space view.

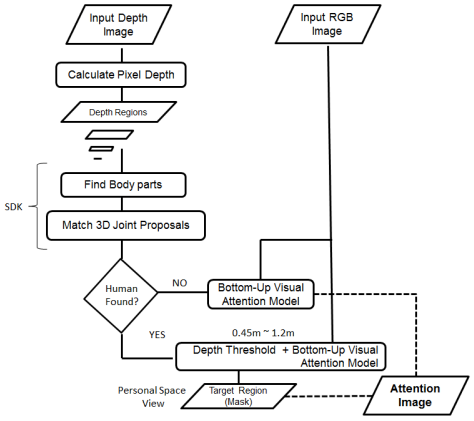

Fig. 9. Attention model that integrates Bottom-Up Visual Attention Model with depth information in order to define focus of attention.

features. In total, six objects (3 salient and 3 non-salient) shown in Fig. 10 were used in the experiment.

The experiment was performed in an ordinary room with no special pre-arranged settings such as plain backgrounds. In fact, our experimental setting contains cluttered background and objects with salient features (i.e. lamp, monitor) that may serve as distractors during the tutoring task.

The experiment was divided into two phases: 1) demonstrating objects with salient features and 2) demonstrating objects with non-salient features. Fig. 11 shows actual experiment images in which the volunteer holds the objects in front of the robot. It can be noticed that non-salient feature objects are difficult to distinguish from 2D image.

In each experimental phase, 3 tutoring tasks with corresponding objects were performed. Each tutoring task lasted 10 seconds and consisted on the following actions:

1) Volunteer stood 1.2m∼1.5m distance from the robot (2 sec).
2) Volunteer performed object demonstration by presenting the object within robot's personal space (6 sec).
3) Volunteer finished demonstration and stepped out of robot's field of view (2 sec).

It is worth mentioning that volunteer was not previously instructed how to perform object demonstration. Volunteer was

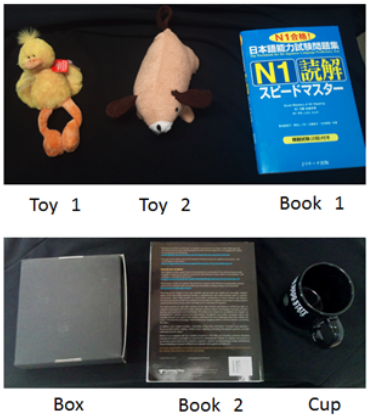

Fig. 10. Experiment objects - upper: objects with salient features, lower: objects with non-salient features.

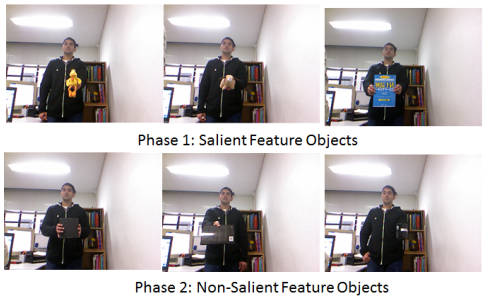

Fig. 11. Phase 1: experiment using objects with salient features. Phase 2: experiment using objects with non-salient features.

free to present the object by lifting the object to desired height, holding the object with one or two hands, move the object in front of the robot, or keep the object still.

B. Evaluation

In order to evaluate which aspects of the demonstration were detected by the attention model, attention locations were classified into four regions: object, tutor's hand, tutor's face, and others (i.e., background objects). Figure 12 (c) shows examples of the classification regions.

Region classification was performed for every frame by examining the center region (20x20 pixel) of the attention image obtained by each attention model. Figure 12 shows the attention region result obtained by the Bottom-Up Attention Model (a) and Depth-base attention model (b). Each center region was classified as object or not depending on whether it was the same color as the object (for salient objects) and by visual inspection (for non-salient objects). Center regions with skin color were categorized as face or hands. Face and hands were then distinguished by the relative position in which hand position is usually lower than the face.

Attention analysis was performed by comparing how often the focus of attention was brought to object, tutor's hand, face or other, using salient and non-salient feature objects.

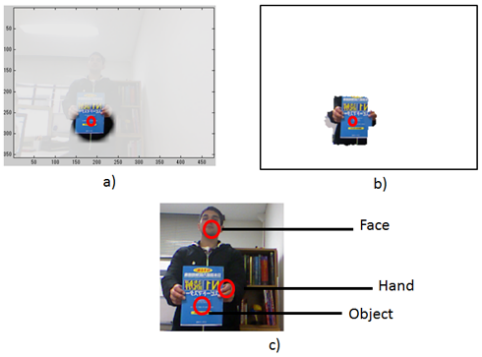

Fig. 12. Example attention regions detected by (a) Bottom-Up Attention Model and (b) Depth-based Attention Model. (c) Classification of attended locations.

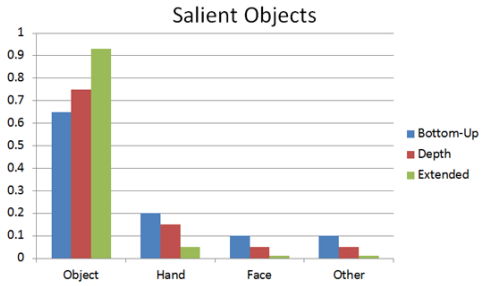

Fig. 13. Results of experiment using objects with salient features. Note that Bottom-Up attention model had a better accuracy in focusing on the demonstrated object compared to the performance of the same model in Fig. 14.

C. Results

Figures 13 and 14 present the proportion of attention of both phases: 1) using salient feature objects and 2) using non-salient feature objects. Each color bar represents the mean proportion of the attention during the three tutoring tasks using a particular type of object: Blue- using only Bottom-Up Visual Attention Model based on Saliency, Red- Depth-based attention model and Green- attention model that combines both approaches.

In Fig. 13 it can be noticed that Bottom-Up attention model had a better accuracy in focusing on the demonstrated object compared to the performance of the same model in Fig. 14. This was mainly due to the fact that salient-feature objects were easier to detect as compared to non-salient feature objects. An interesting point is that Bottom-Up model was able to focus on the object for some short period of time even with non-salient objects. This may be the result of the object movement done by the volunteer during the demonstration. Therefore, we can confirm that motionese is also a very important factor that defines the visual focus of attention.

In Fig. 14, it can also be noticed that the Bottom-Up attention model was highly distracted by the hand, face and background. On the other hand, depth-based attention model alone performed fairly well during the demonstration of salient and non-salient feature objects. This result seems reasonable since most of the times the object was demonstrated within the robot's pre-defined depth threshold distance. However, we

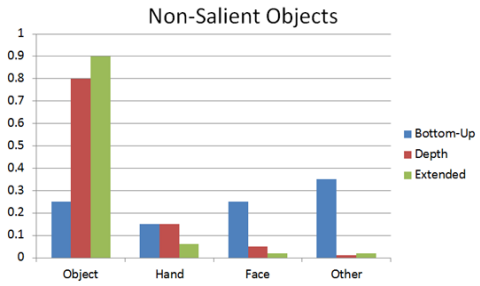

Fig. 14. Results of experiment using objects with non-salient features. Note that Bottom-Up attention model was highly distracted by the hand, face and background. Depth-based attention model alone performed fairly well during the demonstration of salient and non-salient feature objects. Our approach that combines depth information along with the Bottom-Up Attention Model represented by the green bar demonstrates higher attention accuracy located in the demonstrated object.

can notice a small proportion of attention directed to the hands of the volunteer that may have held the object with two hands or with one hand covering part of the object.

Finally, our proposed attention model that uses depth information along with the Bottom-Up Attention Model represented by the green bar demonstrates higher attention accuracy located in the demonstrated object. While the performance does not differ too much from the depth based attention model, the improvement may have been caused by using the Bottom-up attention model to find the salient feature object even when the volunteer did not present the object within the pre-defined depth threshold distance.

V. Conclusion

In this paper we discussed the factors that have an important effect in infants' attention and the way these factors can be taken into consideration to develop robot attention models that simulate infants' cognitive stimuli. We proposed an attention model that uses depth information along with the Bottom-Up Attention Model based on Saliency to increase attention accuracy of objects during a tutoring task when a human tutor is present. Our model can be used for robots to locate the focus of attention in objects that are presented at a close distance, even if objects do not have salient features. Experimental results show that depth information plays an important role for defining the focus of attention of systems that emulate the cognitive development of infants.

VI. Future Work

In future work we will perform experiments with a richer variety of objects, and we will compare saliency performance across multiple volunteer tutors.

Acknowledgment

This work was supported in part by Grant-in-Aid for Scientific Research (C) 23500242.

References

[1] Smith, L.B., Yu, C. and Pereira, A.F. (2007). From the Outside-In: Embodied Attention in Toddlers. In Proceedings of 9th European Conference of Artificial Life (ECAL2007) (pp. 445-454).
[2] C. Breazeal and A. L. Thomaz. "Learning from human teachers with socially guided exploration." In Proceedings of the International Conference on Robots and Automation (ICRA), 2008
[3] A. L. Thomaz and M. Cakmak. "Learning about Objects with Human Teachers." In Proceedings of the International Conference on Human Robot Interaction (HRI), 2009
[4] A. Vogel, K. Raghunathan, D. Jurafsky "Dialog with Robots". In *AAAI* 2005.
[5] Mansur, A. Sakata, K. Rukhsana, T. Kobayashi, Y. Kuno. "Human robot interaction through simple expressions for object recognition". The 17th IEEE International symposium on Robot and Human Interactive Communication. *RO-MAM* 2008.
[6] Y. Nagai and K. J. Rohlfing, "Computational Analysis of Motionese Toward Scaffolding Robot Action Learning," IEEE Transactions on Autonomous Mental Development, vol. 1, no. 1, pp. 44-54, May 2009
[7] Y. Nagai, A. Nakatani, and M. Asada, "How a robot's attention shapes the way people teach," in Proceedings of the 10th International Conference on Epigenetic Robotics, pp. 81-88, November 2010
[8] R.J. brand, D.A. Baldwin, and L.A. Ashburn, "'Evidence for 'motionese':Modifications in mothers' infant-directed action," Developmental Science, vol. 5, no.1, pp.72-83, 2002.
[9] K.J. Rohlfing, J.Fritsch,B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" Adv, Robot., vol.20, no. 10. pp. 1183-1199, 2006.
[10] Y. Nagai and K. J. Rohlfing, "Parental action modification highlighting the goal versus the means," in Proc. IEEE 7th Int. Conf. Develop. Learning, 2008.
[11] Begum, M.; Karray, F.; , "Visual Attention for Robotic Cognition: A Survey," Autonomous Mental Development, IEEE Transactions on , vol.3, no.1, pp.92-105, March 2011
[12] C.Koch and S. Ullman, "Shifts in selective visual attention: Toward the underlying neural circuitry," Human Neurobiol., vol. 4, pp. 219-227, 1985.
[13] R. Bajscy, "Active perception," Proc. IEEE, vol. 76, pp. 996-1005, 1988.
[14] T.Jost, N.Ouerhani, R.Wartburg, R. Muri, H. Hugli, "Contribution of Depth to Visual Attention", Proc. Early Cognitive Vision Workshop, 2004.
[15] T.Jost, N.Ouerhani, R.Wartburg, R. Muri, H. Hugli, "Computing Visual Attention from Scene Depth", Proc. International Conference on Pattern Recognition, pp. 375-378, 2000
[16] M.Z. Aziz and B. Mertsching, "Fast Depth Saliency from Stereo for Region-Based Artificial Visual Attention", in Proc. ACIVS, pp.367-378, 2010.
[17] A. Maki, P. Nordlund, J. O. Eklundh, "A computational model of depth-based attention", In Pattern Recognition, 1996., Proceedings of the 13th International Conference on, Vol. 4 (Aug 1996).
[18] Jamie Shotton, Andrew W. Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, Andrew Blake: Real-time human pose recognition in parts from single depth images. CVPR 2011: 1297-1304
[19] L. Itti and C. Koch, "Computational modeling of visual attention," Nature Rev.: Neurosci., vol. 2, pp. 194-203, 2001.
[20] L. Itti C. Koch and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Analysis and Machine Intelligence. vol 20, no.11, pp.1254-1259, Nov. 1998.
[21] Y.Sun and R. Fisher, "Object-based visual attention for computer vision," Artificial Intelligence vol. 146, pp.77-123, 2003.
[22] S. Frintrop, "VOCUS: A Visual Attention System for Object Detection and goal-directed Search". Heidelberg, Germany: Springer-Verlag, 2006, vol. 3899. LNAI 3-540-32759-2.
[23] Hall, E. T. (1966). *The Hidden Dimension*. New York: Anchor Books.
[24] Microsoft Kinect for Windows SDK BETA from Microsoft Research, http://research.microsoft.com/en-us/um/redmond/projects/kinectsdk
[25] Jonas Ruesch, Manuel Lopes, Alexandre Bernardino, Jonas Hornstein, Jose Santos-Victor, Rolf Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot iCub" In proceeding of: 2008 IEEE International Conference on Robotics and Automation, ICRA 2008, May 19-23, 2008, Pasadena, California, USA.