Can state-of-the-art saliency systems model infant gazing behavior in tutoring situations?

Tadmeri Narayan Vikram, Katrin S. Lohan, Marko Tscherepanow, Katharina J. Rohlfing, and Britta Wrede
Bielefeld University, Applied Informatics Group, CoR-Lab
Email: nvikram@cor-lab.uni-bielefeld.de

I. Introduction

The behavior for a humanoid robot is often modeled in accordance with human behavior. Current research suggests that analyzing infant behavior as a basis for designing the robot behavior can guide us to a natural robot interface. Based on this idea many researchers support saliency systems as a bottom-up inspired way to simulate infant-like gazing behavior. In the field of saliency systems many different approaches have proposed and quantified in terms of speed, quality and other technical issues. But so far, no one compared and quantified them in terms of natural infant tutor interaction.

The question we would like to address in this paper is: Can state-of-the-art saliency systems model infant gazing behavior in tutoring situations? By addressing these issues we want to take a step towards an autonomous robot system, which could be used more natural interaction experiments in future.

II. Methods for Analyzing

We compared 7 different saliency systems with the gazing behavior of infants between the age of 8-11 months. The infant's gazing behavior and the interaction structure of the tutoring situation was manually annotated. Based on this annotation we selected images as input for the saliency systems. The interaction data used in the course of this research is described in the following subsections. The employed saliency systems are briefly described in subsection II-B.

	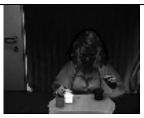	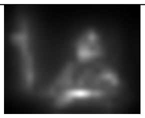	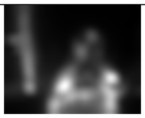	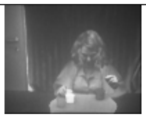	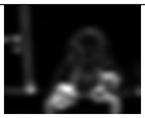	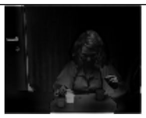	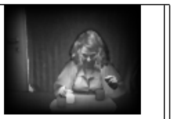
Original Image	Frequencytuned Saliency II-B	Graphbased saliency II-B	Mutliresolution saliency II-B	Random center-surround II-B	Local steering kernel II-B	Symmetry II-B	Random rectangluar regions II-B

TABLE I
An input image and the resultant saliency maps.

A. Parent-child interaction

Out of the well-known Bielefeld-motionese corpus [8], we focus on the youngest participants comprising 12 families of 8 to 11 months old children to analyze their gazing behavior, as the main feedback and controlling capabilities of these infants is based on the gazing behavior [13]. We selected the stacking cups task, because it was often analyzed [7],[12] and the interaction is very detailed analysed. We selected several pictures based on an action segmentation. The pictures represent the beginning and ending point of the stacking cups task. We have chosen this points because at these points the scene is changing.

Setting: Parents were instructed to demonstrate a stacking cups task to an interaction partner. The first interaction partner was their infant, and another adult constituted the second. Fig. 1 illustrates the top-view of the experimental setup, and shows sample image frames of cameras which were set behind the parent and the interaction partner and focused on each of them. The stacking cups task was to sequentially pick up the green (a1), the yellow (a2), and the red (a3) cup and put them into the blue one on the white tray.

B. Saliency Systems

Frequency tuned saliency model [1] is perhaps the simplest of all the existing saliency systems. The absolute difference of each pixel to the image mean is accounted as its saliency value. They recommend to decompose a given input image into L*, a*, b* color space and perform the aforementioned operation on each of the color plane and fuse the results to compute the final saliency map.

Graph-based visual saliency models [4] envisages the input image as a complete graph with each pixel as its nodes. A Weber's [3] law based dissimilarity measure is employed to compute the dissimilarity between any two given pixels. A normalizing function is further employed to compute the final saliency map. Experiments carried out in [4], [10] have shown that the method has high performance in correlating with human eye gaze and also for object segmentation. *The Multi resolution saliency map* [5] is perhaps the most cited saliency system to date. The authors propose a computational model for the theoretical framework presented by Koch and Ulman [6] to model visual attention. An input image is analyzed with opponent color maps, pixels gradient and orientation at different scale spaces. A winner take all (WTA) network fuses these multiple maps into a single saliency map.

Local steering kernel-based saliency [9] is a based on the center-surround paradigm, which is employed to compute the saliency of a selected region. They propose a novel local steering kernel which is employed to compute the similarity between a center region and the surrounding patches. The methodology is inherently robust to image brightness and contrast changes, has very few parameters that requires to be fined-tuned.

Symmetry based Saliency [2] is compute rectangular regions of interest centered on a pixel and compute first order moments of features to calculate the saliency of the given pixel. The methodology is among the most best in terms of programming simplicity. The authors also claim the biological plausibility of the model.

Random center-surround pattern based Saliency [11] is a methodology that employs a biologically plausible dissimilarity metric which is employed to compute the contrast between any two random pixels on the input image. The contrasts are updated and normalized to generate the final saliency map. The said methodology is shown to have state-of the-art performance in the task of salient region detection.

For the *Random rectangular regions of interest based saliency* [10], the same authors as in [11] propose to compute frequent-tuned salient region detection [1], on random rectangular regions of an image for a large number of times and then sum them to compute the final saliency map. Such a formulation is shown to have excellent correlation with human eye-gaze and has good performance for the task of salient regions as shown in their experiments. The methodology also has only two parameters which requires fine tuning, which reduces the implementation complexity.

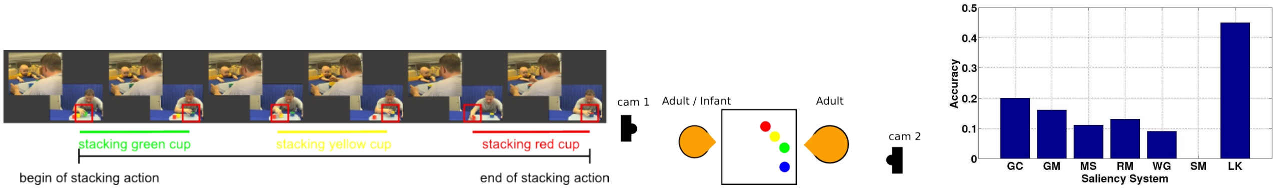

Fig. 1. The first image shows the action segmentation of the stacking cups task, the middle image is giving an insight into the setting of the motionese corpus and the last graphic is representing our results.

III. Data

The beginning and concluding snapshots of 12 motionese videos were cropped, and were parsed into the seven aforementioned saliency systems. A 15 x 15 region centered on the maximally salient point was chosen as the focus of attention. see Table I.

Results: Results are summarized in Fig. 1. The percentages describes the matching accuracy of a given saliency system to the child's eye gazing behavior. It seems that the methodology of Seo and Milanfar [9] has the best matching with the gazing behavior of a 8-11 month old child.

Acknowledgment

Tadmeri Narayan Vikram gratefully acknowledges the financial support from the EU FP7 Marie Curie ITN RobotDoc. Contract No. 235065. Katrin S. Lohan, Katharina Rohlfing, and Britta Wrede gratefully acknowledge the financial support from the FP7 European Project ITALK.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. 2009.

[2] R. Achanta and S. Susstrunk. Saliency detection using maximum symmetric surround. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 2653–2656. IEEE, 2010.

[3] B. Chanda and D.D. Majumder. *Digital image processing and analysis*. PHI Learning Pvt. Ltd., 2004.

[4] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545, 2007.

[5] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[6] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 4(4):219–27, 1985.

[7] K. S. Lohan, A. L. Vollmer, J. Fritsch, K. Rohlfing, and B. Wrede. Which ostensive stimuli can be used for a robot to detect and maintain tutoring situations? IEEE, 2009.

[8] K.J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann. How can multimodal cues from child-directed interaction reduce learning complexity in robots? *Advanced Robotics*, 20(10):1183–1199, 2006.

[9] H.J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12), 2009.

[10] T. N. Vikram, M Tscherepanow, and B. Wrede. A saliency map based on random sub-window means.

[11] T.N. Vikram, M. Tscherepanow, and B. Wrede. A random center surround bottom up visual attention model useful for salient region detection. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 166–173. IEEE.

[12] A. L. Vollmer, K. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K. Rohlfing, and B. Wrede. People modify their tutoring behavior in robot-directed interaction for action learning. In *International Conference on Development and Learning*, volume 8, Shanghai, China, 04/06/2009 2009. IEEE, IEEE.

[13] A. L. Vollmer, K. Pitsch, K. S. Lohan, J. Fritsch, K. Rohlfing, and B. Wrede. Developing feedback: How children of different age contribute to an interaction with adults. In *International Conference on Development and Learning*, 2010.