Berthouze, L., Kaplan, F., Kozima, H., Yano, H., Konczak, J., Metta, G., Nadel, J., Sandini, G., Stojanov, G. and Balkenius, C. (Eds.)
Proceedings of the Fifth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems
Lund University Cognitive Studies, 123. ISBN 91-974741-4-2

Out in the World: What Did The Robot Hear And See?

Lijin Aryananda
MIT CSAIL, 32 Vassar St Rm 380, Cambridge, MA 02139, lijin@csail.mit.edu

1. Introduction

In this paper, we present preliminary results obtained during the early development phase of a humanoid head robot, MERTZ, to explore socially situated learning. Inspired by the infant's learning process, we strongly emphasize the role of experience in the world and social interaction, which are likely to be crucial to the development of human intelligence (Vygotsky, 1978). Given these biases, we propose that the robot must be able to operate in human spaces and time scales, continuously running everyday in different locations, and interact with many passersby. The idea is to move toward a robotic creature that *lives* around people on a regular basis and incrementally learns from its experience instead of a research tool for a specific task that interfaces only with its programmers within short-term testing periods. Starting with a set of *innate* behaviors and drive, i.e. preference for faces etc, we plan to explore the tasks of unsupervised individual recognition and lexical acquisition of socially relevant words.

Developmental approaches to robotics have been explored by many researchers, as surveyed in (Lungarella et al., 2004). The importance of social interaction for the robot learning framework was proposed in (Dautenhahn, 1995) and have been implemented in many platforms (Fong et al., 2003). The target task of word learning in robots through interaction with humans have also been explored in (Roy et al., 2002, Steels et al., 2001).

In (Aryananda et al., 2004), we report on design steps and issues involving the challenge of achieving reliable continuous operation. In this paper, we address the equally challenging objective of generating social feedback from humans which would be crucial to the robot's learning process. In particular, we analyze the feasibility for learning to recognize individuals and acquire lexicon of relevant words through social interaction. To this end, we have developed the perceptual and behavior systems to engage in simple visual and verbal interaction. The robot tracks salient visual targets (faces and bright objects) and tries to encourage verbal interaction by mimicking phoneme sequences extracted from speech input.

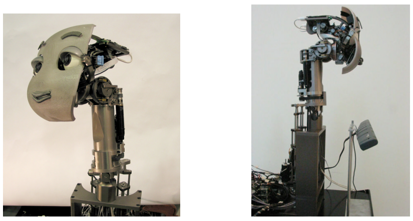


Figure 1. MERTZ, an active-vision humanoid head robot, mounted on a portable platform.

We conducted an experiment where the robot ran for 7 hours/day for 5 days at different locations on the 1st floor of the laboratory building (Stata Center), to address the following questions:

Will people interact with the robot enough to acquire a set of training face images for further recognition? Are people going to speak to the robot? Will they use enough one-word utterances to avoid the difficult word segmentation task? Will there be enough word repeatability for the robot to acquire a lexicon of socially relevant words?

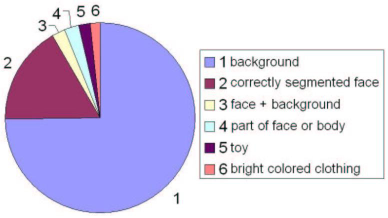
2. Robotic System

MERTZ is an active-vision head robot with thirteen degrees of freedom (see Figure 1), two digital cameras, and a voice array desk microphone. The robot's visual system[1] is equipped with detectors for skin and saturated color, motion, and face (Viola et al., 2001). These detectors are complemented with a KLT-based face tracker (Shi et al., 1994) and a color-histogram tracker. The audio system consists of a phoneme recognizer (Mosur) and the DECtalk speech synthesizer. More implementation details are described in (Aryananda et al., 2004)

3. Experimental Results

What did the robot see? We labelled the tracker's output from a sequence of 14186 frames collected during a 4-hour period on day 2 (see Figure 2 Top) For 16.9% of the time, the robot tracked correctly segmented faces. For a small part of the time, the robot tracked partially segmented faces

[1] The visual system is implemented using YARP (Http://sourceforge.net/projects/yarp0)



Day	Face	Non-face	% accuracy
1	37749	6350	87.82
2	26311	3649	84.26
3	20199	3772	85.6
4	10167	6683	60.34

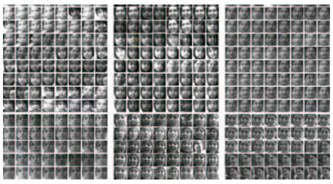


Figure 2. Top: The characteristic of speech input received by the robot on each day of the experiment. Bottom: The face detector's accuracy over 114,880 faces detected in 4 days and a sample set of the 94,426 correctly detected faces.

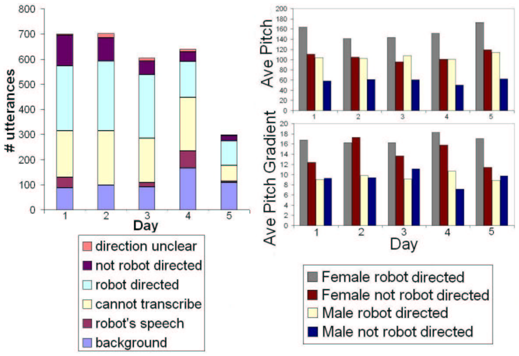
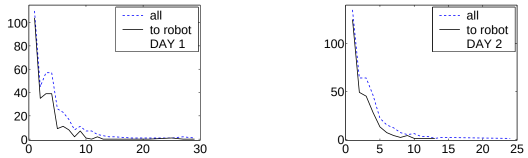


Figure 3. Left: The characteristic of speech input received by the robot on each day of the experiment. Right: Pitch average and gradient values from robot and non-robot directed speech.

and bright objects. We also collected every frame of detected faces throughout the experiment, which amount to 114,880 images from at least 600 individuals. Figure 2 Bottom shows the detector's accuracy during each day, excluding one due to file loss. These results suggest that it is feasible to acquire a significant set of face images because people do interact with the robot closely and for long enough durations.

What did the robot hear? The robot received over 1000 utterances (segmented using energy threshold) per day. We transcribed 3027 audio samples recorded from at least 200 speakers, taken from a few continuous sequences of speech input spanning several hours each day. As shown in Figure 3 Left, approximately 37% of the total utterances are intelligible robot directed speech. Figure 4 Top shows that one-word utterances make up 38% of all intelligible speech and 38.6% of robot directed speech. Approximately 83.2% of all intelligible speech and 87.8% of robot directed speech contain less than 5 words. Figure 4 Bottom illustrates the top fifteen most frequently said words during each experiment day. Figure 3 Right illustrates that average pitch of female and male speakers is higher when speaking to the robot versus other people. The pitch gradient data suggests that female speakers produce more modulated pitch contours when speaking to the robot. These results suggest that people do verbally interact with the robot. The frequency of one-word utterances seems to be high enough to provide the robot with a starting point for unsupervised lexical acquisition. Lastly, a set of common words tend to be repeated throughout the experiment despite the large number of speakers and minimal constraints on the human-robot interaction.



	15 most frequent words
Day 1	it/it's,you,hello,I,to,is,what,hi,are,the,a,mertz,here,your,this
Day 2	you,hello,it/it's,what,I,hi,yeah,are,the,oh,to,is,mertz,a,your
Day 3	hello,it/it's,you,hi,mertz,bye,to,robot,are,the,I,what,hey,how,is
Day 4	you,hello,it/it's,what,hi,mertz,are,I,bye,this,here,how,is,robot,to
Day 5	you,hello,it/it's,hi,I,what,are,oh,how, say,the,a,at,can,is

Figure 4. Top: Number of words/utterance collected on day 1&2. Bottom: The top 15 most frequent words

References

L. Aryananda, J. Weber. Mertz: A Quest for a Robust and Scalable Active Vision Humanoid Head Robot. *IEEE-RAS/RSJ International Conference on Humanoid Robots*, 2004.

K. Dautenhahn. Getting to Know Each Other-Artificial Social Intelligence for Autonomous Robots. *Robotics and Autonomous Systems*, 16:333-356, 1995.

T. Fong, I. Nourbakhsh, and K. Dautenhahn. A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems*, 42:143-166, 2003.

M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental Robotics: A Survey. *Connection Science*, vol.00 no.0:1-40, 2004.

R. Mosur. Http://cmusphinx.sourceforge.net/sphinx2.

D. Roy and A. Pentland. Learning Words From Sights and Sounds: A Computation Model. *Cognitive Science*, 26(1):113-146, 2002.

J. Shi and C. Tomasi. Good features to track. *IEEE Conference on Computer Vision and Pattern Recognition*, pp 593-600, 1994.

L. Steels and F. Kaplan. Aibo's First words. The Social Learning of Language and Meaning. *Evolution of Communication*, 4(1), 2001.

P. Viola and M. Jones. Robust Real-time Object Detection. *Technical Report Series, CRL2001/01. Cambridge Research Laboratory*, 2001.

L. Vygotsky. Mind in Society. *Cambridge, MA: Harvard Univ Press*, 1978.

Segmentation Stability: a Key Component for Joint Attention

Jean-Christophe Baillie* **Matthieu Nottale***
*ENSTA / Cognitive Robotics Theme
Laboratory of Electronics and Computer Engineering,
32, boulevard Victor, 75739 Paris Cedex 15 France

Abstract

It is now well established that joint attention is a key capability for socially interacting robots (Brooks et al., 1999, Kaplan and Hafner, 2004, Scassellati, 1999, Itti, 2003). It is also a key component for epigenetic robotic applications in general. This subject has been widely discussed and we present here one specific technical improvement for joint attention which relies on image segmentation.

In the new Talking Robots (Baillie, 2004) experiment that we have started, following a successful reimplementation of the Sony's Talking Heads (Steels, 1998) experiment on Aibo ERS7, we try to have two robots interacting to evolve a shared repertoire of synchronized behaviors, or "games".

This leads to a dynamic version of the Talking Heads, where the interaction protocol, or language game, is not predefined in the agents. As usually with experiments involving symbol grounding and social interaction, and for these experiments in particular, it is essential that the two robots can establish joint attention and share a common representation of their surrounding environment, while they are looking at the same scene. Many techniques can be used to achieve this stable shared representation. We have chosen to focus on segmentation-based algorithms, which provide interesting shapes together with vectors of features that can be easily extracted and proved useful in the context of our experiment.

However, when using segmentation algorithms, a slight change in the viewpoint, or even the residual camera noise, is enough to significantly change the result of the image partition, possibly leading to completely different perceptions of the same scene.

We have developed an original measure to assess the stability of a segmentation algorithm and we have used it on a set of algorithms to automatically determine the most stable partition for a given scene. This approach is different from classical methods used to estimate the quality of a segmentation algorithm, where the result of the algorithm is compared to an ideal perfect segmentation done by hand. In our approach, the measure is done automatically, involves only stability considerations and could lead to interesting improvements whenever joint attention using segmentation is required.

We quickly present in the poster the background of the Talking Robots experiment and why joint attention and image segmentation stability is an important issue for us. We introduce then two stability measures and show some results on natural scenes from our experiments in the lab and test scenes used to control image parameters. The influence of several image characterizations (noise, number of objects, luminosity,...) is carefully reviewed. An example (fig.1) is given below, showing the influence of noise on typical images.

Using the fact that a given algorithm can be ranked according to its stability score, which is calculated online assuming that the scene itself is static, a general method of algorithm switching is introduced and used in the experiment with different kind of algorithms: region growing, recursive histogram splitting, CSC (Priese and Rehrmann, 1993) and split & merge (CIS). We show how this method significantly improves the convergence speed in the experiment and conclude on the generality of our approach to facilitate certain aspects of joint attention, when it relies on segmentation.

References

Baillie (2004). Grounding symbols in perception with two interacting autonomous robots. *Proceedings of the 4th International Workshop on Epigenetic Robotics*, 117.

Brooks, R. A., Breazeal, C., Marjanovic, M., Scassellati, B., and Williamson, M. M. (1999). The cog project: Building a humanoid robot. *Lecture Notes in Computer Science*, 1562:52–87.

Itti (2003). Visual attention. In Arbib, M. A., (Ed.), *The Handbook of Brain Theory and Neural Networks, 2nd Ed.*, pages 1196–1201. MIT Press.

Kaplan, F. and Hafner, V. (2004). The challenges of joint attention. *In: Proceedings of the 4th International Workshop on Epigenetic Robotics.*

Priese, L. and Rehrmann, V. (1993). A fast hybrid color segmentation method. In *DAGM-Symposium*, pages 297–304.

Scassellati, B. (1999). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. *Lecture Notes in Computer Science*, 1562:176–195.

Steels, L. (1998). The origin of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103:133–156.

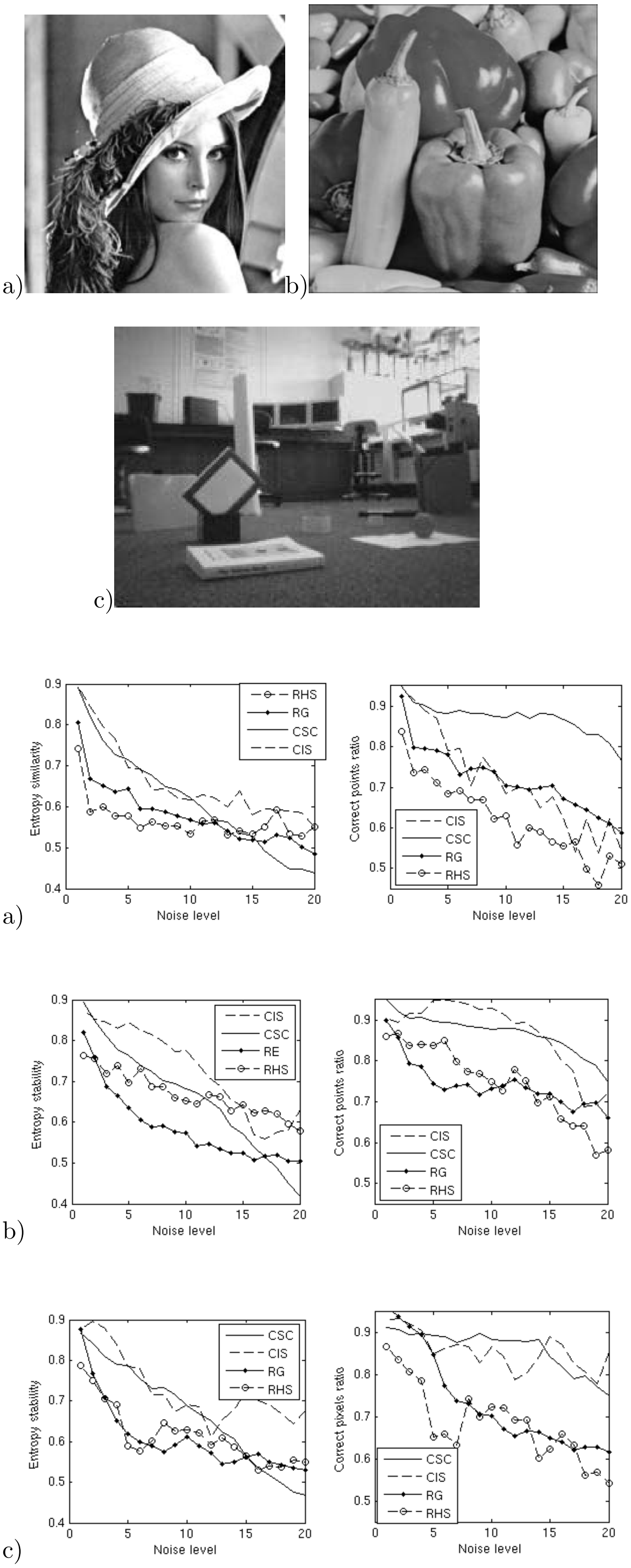


Figure 1: Some stability measure results for image Lena, peppers and an image of our lab, against varying noise level (standard deviation of pixel graylevel). Most stable images have a stability of 1.