In Berthouze, L., Kozima, H., Prince, C. G., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C. (Eds.)
Proceedings of the Fourth International Workshop on Epigenetic Robotics
Lund University Cognitive Studies, 117, ISBN 91-974741-3-4

Grounding Symbols in Perception with two Interacting Autonomous Robots

Jean-Christophe Baillie
ENSTA / Electronics and Computer Engineering Lab.
Paris 75015 France
baillie@ensta.fr

Abstract

Grounding symbolic representations in perception is a key and difficult issue for artificial intelligence. The ”Talking Heads” experiment (Steels and Kaplan, 2002) explores an interesting coupling between grounding and social learning of language. In the first version of this experiment, two cameras were interacting in a simplified visual environment made of colored shapes on a white board and they developed a shared, grounded lexicon. We present here the beginning of a new experiment which is an extension of the original one with two autonomous robots instead of two cameras and a complex and unconstrained visual environment. We review the difficulties raised specifically by the embodiment of the agents and propose some directions to address these questions[1].

1. Introduction

Grounding symbolic representations into perception is a key and difficult issue for artificial intelligence (Harnad, 1990, Ziemke, 1997, Brooks, 1990, Siskind, 1995, Roy, 2002). Including social interactions and, more specifically, language acquisition and development in this context has proven to be a fruitful orientation. One of the recent and successful attempts in this direction is the ”Talking Heads” experiment (Steels and Kaplan, 2002, Steels, 2001, Steels, 1998). This experiment involves two cameras interacting in a simplified visual environment made of colored shapes on a white board. Because of this simplified environment, the range of lexical items that can be grounded is limited to simple notions like size, position on the board or simple color categories. With the recent development of relatively cheap and powerful robotic platforms (see Sony, Honda or Fujitsu, among others) research on symbol grounding can move from simulation or simple environments to complex natural environments and embodied systems. Following this trend, we present here an attempt to reproduce the initial Talking Heads experiment with autonomous robots (Aibo ERS7) evolving in an unconstrained visual environment, instead of simple cameras. The goal is to reinforce the validity of the first results of the Talking Heads experiment in showing that the proposed mechanism for lexicon acquisition stands in the case of a noisy, complex environment. Previous attempts to realize this extension of the Talking Heads experiment have been conducted with a robot/human interaction (Steels, 2001, Steels and Kaplan, 2001) or with robot/robot interaction with simple visual perception (Vogt, 2000), but to our knowledge, none has been conducted with a complex robot/robot interaction in an unconstrained visual environment, which is what we investigate.

Language games are the theoretical basis supporting the Talking Heads experiment and they have been thoroughly described in (Wittgenstein, 1967, Steels and Kaplan, 2002). We present how language games can be implemented in an autonomous robotic device and what are the key issues associated. Clearly, implementing the Talking Heads in robots is not only a matter of more complex vision algorithms but involves difficult questions in control, behavior, learning and categorization.

2. Embodied ”Talking Heads” experiment

2.1 Description

The purpose of our experiment is to reproduce the guessing game (Steels, 2001) with two important differences from the first Talking Heads experiment:

1. The agents are no longer simple cameras but autonomous robots (Aibo ERS7).

2. The environment is no longer a simple set of shapes on a white board but an unconstrained image from the lab where the robots are.

Why is this extension of the experiment important to do? First, it is likely that to develop more complex grounded symbols, it is also necessary that the perceived

[1] This work is supported by a three years funding from the French DGA

environment becomes more complex. Second, the problem is extremely difficult and raises a number of issues that go beyond the language acquisition problem. From this point of view, this experiment can be seen as an integrating application for many research domains from control, perception, categorization, and robotics learning.

Below is a review of the different problems raised by the embodiment of the Talking Heads experiment and possible solutions that we will examine. We will always try to prefer solutions that could be conceivably extended to a robot/human interaction since this is a natural future work for the Talking Heads experiment.

2.2 Requesting attention

In the original experiments, the two cameras where transmitting synchronization information to each other by TCP/IP. There was no need for such a thing as ”requesting each other's attention” since the agents were precisely positioned and the start signal was given to each of them at the same time.

With autonomous robots, we cannot make simplifying assumptions on the position or internal state of the robot. The first stage of the experiment is to establish a visual contact. For practical reasons, when the experiment starts the robots should be at a reasonable distance from each other, and no occlusion should prevent one robot from seeing his partner.

The detection of another robot in the image is closely related to researches on the more general problem of faces or objects detection (Yang et al., 2002). Multi resolution appearance-based methods, using neural networks (Rowley et al., 1998) or Bayes classifiers (Schneiderman and Kanade, 1998), have shown good performances for face or car detection and can be used in the context of detecting an ERS7 Aibo robot.

In our case, the problem is simplified since the hearer knows that he is searched by the speaker and it can actively help him. We let the hearer oscillates its head slightly around the zero position, while the speaker is doing an image substraction to detect moving elements (see Fig. 1). Since we assume the head is the only moving part in the image, the result is an easy to extract pattern close to the edges of the head. By matching this pattern to a previously known set of patterns, it is possible to estimate the orientation of the head, by taking the orientation of the closest known pattern. Since the head is in the zero position, this method gives also the orientation of the body. This method has been implemented and gives encouraging results.

After the speaker has established this visual contact, it must position itself at a predefined short distance from the hearer. The apparent size of the head pattern in the image is used as a rough distance measure. This positioning stage is necessary to prepare the robots for the attention sharing phase, for which an approximatively common visual context is required.

Once the robot is properly positioned towards its partner, it has to require its partner's attention in order to start the game. The signal used for requesting attention is a simple beep repeated at one second interval. In fact, any appropriate sound could be used here but beeps are commonly used for low-noise communication between Aibo robots and work with very high accuracy. The first robot that request the other one's attention will be the speaker later. When the beep repetition stops, both robots knows that they are ready for a game.

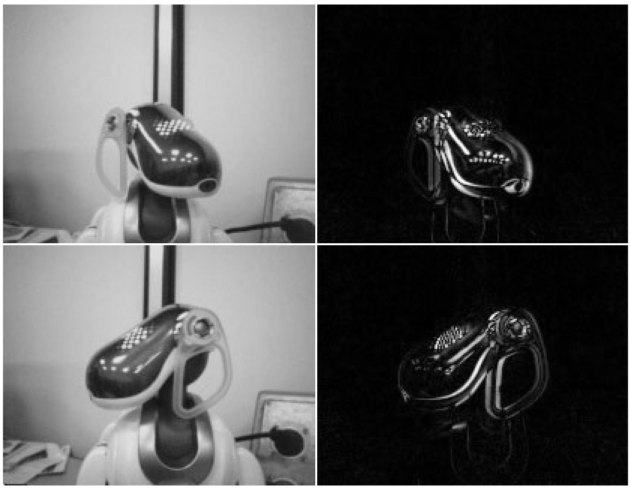

Figure 1: Detecting a robot's head orientation using slight movements around a central position.

2.3 Sharing attention

At the end of the attention request phase, both robots are ready to start the game, one is the speaker and one is the hearer and they are positioned at a short distance from each other. The purpose of the attention sharing process that follows is to ensure that the visual context during the experiment will not be too different from one robot to the other and, more specifically, that the topic chosen by the speaker is visually accessible to the hearer.

The speaker chooses a sight direction inside two cones centered on the axis orthogonal to the line connecting the robots. This ensures that at the end of the attention sharing process, both the speaker and the hearer will share a view where no robot is visible. This is a necessary condition, since the hearer cannot see itself, and thus cannot share a visual context where it might be visible.

The next question is how can one robot detect where the other one is looking? We have simplified the problem by using a small laser pointer attached to the robot's head as a pointing device. The red dot in the image is very specific and very bright which helps to extract it from the rest of the image. Previous studies like (Kirstein, 1998) report successful use of a laser pointer to designate points of interest in unconstrained images. We also improve the efficiency of the detection by controlling the laser activation, instead of having it permanently switched on. By blinking the red dot and using image differentiation, the localization is easier and very accurate (almost no failures, except when the laser points to a reflecting or metallic surface).

To insure that the two robots will look in the same direction, the hearer tries to locate the speaker's pointer and, once both robots are ready (sound synchronization), they both center their view on the laser pointer. It is not enough for the hearer to center its view on the pointer, the speaker must do it too since there is no a priori certitude that the red dot will be in the center of the vision field for the speaker.

In this direction, the visual context of the hearer is supposed to be consistent with the one of the speaker. The previous Talking Heads experiment has shown that small and reasonable context differences do not prevent the convergence of the lexicon.

2.4 Perceptual segmentation

We assume at this stage that the speaker and the hearer are both looking at the same point and that they are close enough to each other, so that they see approximatively the same context. We also assume that the objects they see are far enough to avoid any occlusion or perspective issues and that the image is still.

The next stage is for the speaker to select a topic among the context. In the Talking Heads experiment, the context was described as a set of objects visible on the white board. A simple color segmentation gave a one-on-one stable mapping between the real world objects and the segmented regions. In our case, this is no longer true and we cannot expect any relationship between the real world objects and the segmented regions. For this reason, we will not speak of ”objects” but simply of ”regions”.

The speaker is running a set of various segmentation algorithms on the image, using color, brightness, texture or saturation as aggregating criteria. For each region, a vector of features associated to predefined sensory channels is calculated including average hue, saturation, brightness, size, orientation and texture. The speaker chooses one region as the topic.

The main problem here is to ensure a stable enough segmentation process so that the speaker and the hearer end up with a comparable set of region maps, given a similar view point. This is far from guaranteed but some segmentation algorithms, like the CSC algorithm (Rehrmann and Priese, 1998), are known to give more stable results. We have started a study of the stability of several algorithms and parameters by performing repeated segmentation on a video flow of a still scene and measuring the overlapping of the resulting region maps.

2.5 Categorization

At this point, the speaker has chosen a topic and the associated sensory channels, based on saliency (more details on saliency in (Steels, 1998)). It will try to categorize the topic according to the vector of feature values on these channels. The categorization mechanism we will use is the same as the one of the original Talking Heads, using rescaling and discrimination trees (Steels, 2001). If the speaker fails to categorize the topic apart from the other regions of the context, a discrimination game takes place, the speaker refines the discrimination trees and select another topic.

The discrimination tree is the product of a recursive binary subdivision of the sensory channel space[2] into equal subspaces, as described by Steels. More elaborate subdivision mechanisms could be used taking into account the distribution of the feature values in the sensory space, leading to methods inspired from clustering techniques. It is not clear however that such a refinement would lead to better qualitative results and it should be investigated in further work. The sensory space could also be multidimensional, however once again the benefit of this complexity is not guaranteed.

2.6 Speech recognition

Once the topic has been categorized and a meaning is selected, the speaker choose the most successful word association or creates a new word and utters the corresponding succession of phonemes. We will use a predefined set of ten to twenty phonemes to build new words. These phonemes will be selected according to two criteria:

1. They should be speakable by humans. This is to prepare future extension of the experiment to a robot/human interaction.

2. The inter-phoneme distance is maximized in order to ease the recognition phase.

Standard speech recognition or pattern recognition techniques can be used here by the hearer to segment the audio stream into a set of phonemes. The task is simplified because the phonemes are known in advance and, in the case of a robot/robot interaction, they are always identical to the predefined patterns. We also make the assumption that the experiment is taking place in a quiet and non resonant environment to reduce noise.

2.7 Pointing mechanism

One essential requirement is the capacity of the robots to point to the topic. In the guessing game, the hearer points to the guessed topic and, in case of failure, the speaker points to the correct topic.

The natural way of pointing would be to point a ”finger” in the direction of the topic and use 3D information from stereo vision to follow the line of interest until it meets an object and then map this object back onto the vision field. In our case, however, it is impossible because Aibo has only one camera and cannot perform stereovision. Furthermore, the leg of the Aibo can be used as a

[2] After normalization, this space is the segment [0,1].

pointing device, but it is clearly not very precise. In any case, the whole process is very noisy and would probably not lead to accurate enough results.

Once again, for the sake of simplicity, we will use here our laser pointer mechanism which has proven to be very accurate. The ”designating” robot points to the topic with its laser and the other robot searches for the blinking red dot to locate the region of interest that is supposed to be the topic.

3. Implementation issues: URBI

The robots can be programmed directly using OPENR objects running on the Aibo. However, OPENR, which is the official Sony Software Development Kit for Aibo, is subject to changes from one generation to the next and is considered by many as being too low-level for complex AI programming (see for example the Tekkotsu layer at CMU, www.tekkotsu.org). For this reason, we have developed URBI, an Universal Robotic Body Interface which works with a client/server architecture to control the robot with simple commands to set/get joint values or camera, speaker and microphone data. There is an URBI server running on the robot and the applications running on remote machines are using URBI clients (via a C++ library) to communicate with the robot.

URBI is very efficient in term of speed and response time. Also, by writing a specific URBI server, this solution allows us to use our work with any kind of robot, including future humanoid or pioneer robots. Full details and downloads of URBI can be found at http://urbi.sourceforge.net.

The final release of URBI will include a complete URBI language with IF, WHILE, FOR, LOOP commands, function and variable definitions.

4. Conclusion

We have presented here our project to reconduct the Talking Heads experiment in the context of two autonomous Aibo robots interacting in an unconstrained visual environment. The task is raising several difficult problems that have been rightfully simplified in the first experiment, among them: requesting attention, sharing attention, stable perceptual segmentation, categorization, speech recognition and reliable pointing mechanisms. Each of these problems is currently worked on in our lab and plausible or existing solutions have been proposed. We also work on a general architecture to assemble the solutions into a functional framework.

Extensions of this work include the ambitious goal of implementing a robot/human interaction and capabilities to grasp object using arm equipped pioneer robots or future humanoid robots. Another promising research direction followed by other labs is to increase the complexity of the notions that can be grounded, moving from lexicon evolution to grammar evolution. We hope the framework we develop here will also prove useful in this context.

References

Brooks (1990). Elephants don’t play chess. *Robotics and Autonomous Systems*, 6(1&2):3–15.

Harnad (1990). The symbol grounding problem. *Physica D 42*, pages 335–346.

Kirstein (1998). Interaction with a projection screen using a cameratracked laser pointer. *Proceedings of the International Conference on Multimedia Modeling (MMM 98), IEEE Cmputer Society Press.*

Rehrmann and Priese (1998). Fast and robust segmentation of natural color scenes. In *ACCV (1)*, pages 598–606.

Rowley, Baluja, and Kanade (1998). Neural network-based face detection. *IEEE Trans. PAMI*, 20(1):23–38.

Roy (2002). Learning words from sights and sounds. a computational model. *Cognitive Science*, 26(1):113–146.

Schneiderman and Kanade (1998). Probabilistic modeling of local appearance and spacial relationships for object recognition. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 45–51.

Siskind (1995). Grounding language in perception. *Artificial Intelligence Review*, 8:371–391.

Steels (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103(1-2):133–156.

Steels (2001). Language games for autonomous robots. *IEEE Intelligent Systems*, pages 16–22.

Steels and Kaplan (2001). Aibo’s first words : The social learning of language and meaning. *Evolution of Communication*, 4(1).

Steels and Kaplan (2002). Bootstrapping grounded word semantics. In Briscoe, T., (Ed.), *Linguistic evolution through language acquisition: formal and computational models*, chapter 3, pages 53–73. Cambridge University Press, Cambridge.

Vogt (2000). Bootstrapping grounded symbols by minimal autonomous robots. *Evolution of Communication*, 4(1):89–118.

Wittgenstein (1967). Philosophische untersuchungen, suhrkamp, frankfurt.

Yang, Kriegman, and Ahuja (2002). Detecting faces in images: A survey. *IEEE PAMI*, 24(1):34–58.

Ziemke (1997). Rethinking grounding. *In Austrian Society for Cognitive Science, Proceedings of New Trends in Cognitive Science - Does Representation need Reality, Vienna.*