Evaluation Metrics for an Experience-based Mobile Artificial Cognitive System*

Liwei Zhang[1], Sebastian Rockel[1], Alessandro Saffiotti[2], Federico Pecora[2],
Lothar Hotz[3], Zhenli Lu[4], Denis Klimentjew[1], Jianwei Zhang[1]

***Abstract*— In this paper, an experience based mobile artificial cognitive system architecture is briefly described and adopted by a PR2 service robot for the purpose of carrying out tasks within the EU-FP7 funded project RACE. To measure the benefit of learning from experience to improve the robustness of the robot's behavior, an FIM (Fitness to Ideal Model) and a DLen (Description Length) based evaluation approach has been developed.**

I. INTRODUCTION

The main goal of RACE is to develop a framework and methods for learning from experiences in order to facilitate an cognitive intelligent system. To achieve this goal, experiences are recorded as semantic spatio-temporal structures connecting high-level representations, including tasks and behaviors, via their constituents at lower levels down to the sensory and actuator level. In this way, experiences provide a detailed account of how the robot has achieved past goals or how it has failed, and what sensory events have accompanied the activities.

To measure success for a given task in a given scenario, we use an approach inspired by model-based validation techniques [1]; namely, we measure the compliance of the actual robot's behavior to the intended ideal behavior for that task in that scenario. Fig. 1 graphically illustrates this principle: the trace of a given execution of the RACE system is compared against a specification of what the ideal behavior should be, resulting in a "Fitness to Ideal Model" (FIM) measure.

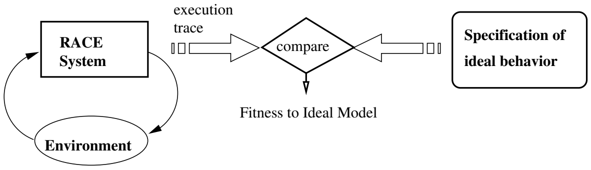

Fig. 1. Principle of evaluation in RACE: the system's behavior is compared to a model of the ideal behavior for the specific scenario.

Discrepancies between the observed behavior and the ideal behavior can originate from errors of four different types: **conceptual**, **perceptual**, **navigation** and/or **localization** and **manipulation errors**. Conceptual errors arise from discrepancies between the knowledge used by the robot and the one encoded in the specification of the ideal behavior. We call these discrepancies *inconsistencies*. Specifically, inconsistencies can be of four types: **temporal**, **spatial**, **taxonomical** and **compositional inconsistencies**. Together, the four above define the FIM metric:

$$\text{FIM} = \sum_{i \in \{t,s,x,c\}} p_i \tag{1}$$

In addition to estimating the effectiveness of learned knowledge by counting the number of inconsistencies, we are also interested in measuring the `Description Length` (DLen [2]) of the instructions that should be given to the robot to achieve a goal. Successful behavior following shorter instruction descriptions is indicative for the effectiveness of the learned knowledge.

II. SCENARIO SET-UP AND EXPERIMENTS

In this work, two demonstrations named "ServeACoffee" and "ClearTable" have been defined and performed on the physical PR2 platform in a restaurant environment. The results are presented and evaluated with respect to the metrics defined and described in [3], [4].

A. ServeACoffee Demonstration

Scenario A: The robot knows the approximate area (pae) of mug1 on counter1, the position of table1, the approx. area of guest1 west of table1, and the areas for manipulation, sitting and placing. The user successively instructs the robot to move to counter1, grasp mug1, move to the manipulation area (mas1) south of table1, and place mug1 at the placing area (pawr1) west of table1. The robot is told that this is a "ServeGuest" activity.

Scenario B: The same as Scenario A, except a new guest (guest2) is sitting east of table1 and the robot is instructed to move to the north of table1 and place mug1 at the east of table1. Again, the robot is told that this is a "ServeGuest" activity.

Scenario C: Guest3 is sitting south of table2 and the robot is simply instructed: Do a "ServeGuest" to guest3.

*This work was supported by the EC Seventh Framework Program theme FP7-ICT-2011-7, grant agreement no. 287752.
[1]L. Zhang, S. Rockel, D. Klimentjew, J. Zhang are with the Department of Informatics, University of Hamburg, Germany {lzhang, rockel,klimentj,zhang}@informatik.uni-hamburg.de
[2]A. Saffiotti, F. Pecora are with the AASS Cognitive Robotic Systems Lab, Örebro University, Sweden {asaffio,fpa}@aass.oru.se
[3]L. Hotz is with the HITeC e.V., University of Hamburg, Germany hotz@informatik.uni-hamburg.de
[4]Z. Lu is with the Institute of Electronics and Telematics Engineering of Aveiro, University of Aveiro, Portugal zhenli.lu@ua.pt

B. Experimental Results

Let $V0$ be the nominal (ideal) condition of the demonstrator (as described in the scenario A schedule). In $V1$ (as described in scenarios B and C), the guest sits on the opposite side of the table (or leaves the sitting area), other than specified in the planning domain. The robot again places the mug in the same area as before (which is now in front of an empty seat). This is classified as perception error and compositional inconsistency.

Here we set weight $\tau_{(\cdot)} = 1$. The initial value of the four types of inconsistency is assigned to be 0. This results in:

$$\begin{aligned} FIM(V0) &= 0 \\ FIM(V1) &= \#\text{spatial\_inconsistencies} \\ &\quad + \#\text{compositional\_inconsistencies} = 2 \end{aligned} \tag{2}$$

Fig. 2 and 3 show the statistical results of the different kinds of errors occurring in the three scenarios. The errors are judged by a human judge during the experiments.

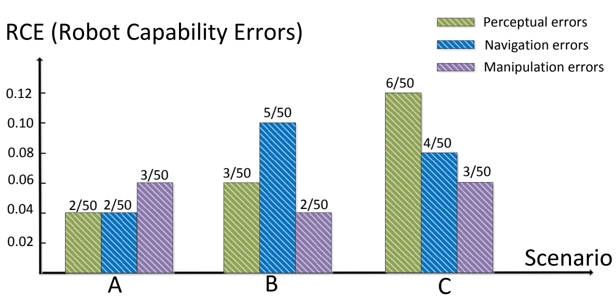

Fig. 2. RCE (Robot Capability Errors) in the ServeACoffee scenario

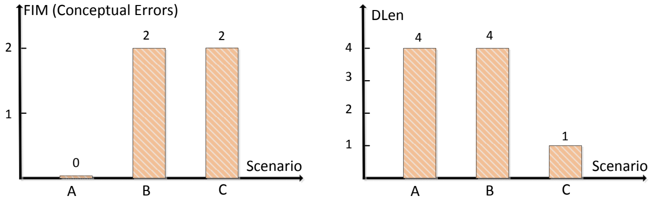

Fig. 3. Conceptual errors (left) and the Description Length of the instructions in ServeACoffee (right)

To measure the Description Length (DLen) of the instructions given to the robot, step by step instructions are provided to the robot. In scenarios A and B, a set of instructions were provided. In the following instruction list of Scenario A, each `achieve` command specifies a sub-task to be carried out by the robot and represents an instruction. The last `teach` command is the instruction to teach a new concept.

1) achieve drive_robot_Task preManipulationAreaEastCounter1
2) achieve grasp_object_w_arm_Task mug1 rightArm1
3) achieve drive_robot_Task preManipulationAreaSouthTable1
4) achieve put_object_Task mug1 placingAreaWestRightTable1
5) teach_task ServeACoffee guest1

In scenario B, similar instructions were provided by the user. In Scenario C, only a single `achieve` instruction is provided as follows. Now the robot can execute the "ServeACoffee" task with a shorter instruction set:

1) achieve serve_coffee_to_guest_Task guest3

Fig. 4 illustrates the relationship between FIM and DLen. The experimental (restaurant) environment is shown in Fig. 5. The scenarios might be executed in the physical or the simulated environment, as indicated by the figures.

III. CONCLUSIONS

The proposed artificial cognitive system has been evaluated with the defined metrics and the results presented. The data obtained indicates an improvement in the robot's knowledge and behavior. Thus the newly introduced metrics are appropriate to evaluate such a system. The initial assumption of an FIM and DLen co-relation is supported by the evaluation results.

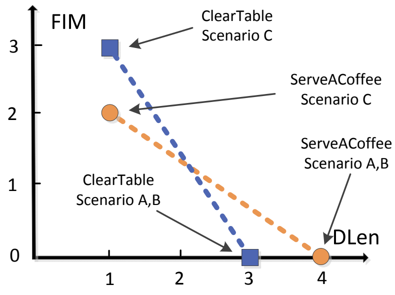

Fig. 4. The relationship between FIM and DLen

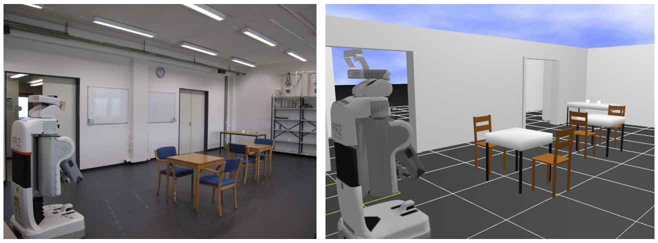

Fig. 5. The restaurant environment in a typical start condition: the robot waits for a guest and to be instructed

REFERENCES

[1] P. S. Kaliappan. Model-based verification techniques: State of the art. Technical report, Brandenburg University of Technology, 2008.

[2] Peter Grünwald. A tutorial introduction to the minimum description length principle. *CoRR*, math.ST/0406077, 2004.

[3] Liwei Zhang, Sebastian Rockel, Federico Pecora, Luís Seabra Lopes, Alessandro Saffiotti, and Bernd Neumann. Deliverable d5.1 - evaluation infrastructure. Technical report, European Commission - Information and Communication Technologies - Seventh Framework Programme, November 2012.

[4] Liwei Zhang and Sebastian Rockel. Deliverable d5.2 - year-1 demonstrator. Technical report, European Commission - Information and Communication Technologies - Seventh Framework Programme, January 2013.