Robot Competitions Kick Innovation in Cognitive Systems and Robotics

RoCKIn

Benchmarking HRI in RoCKIn competitions

Francesco Amigoni
(Politecnico di Milano)

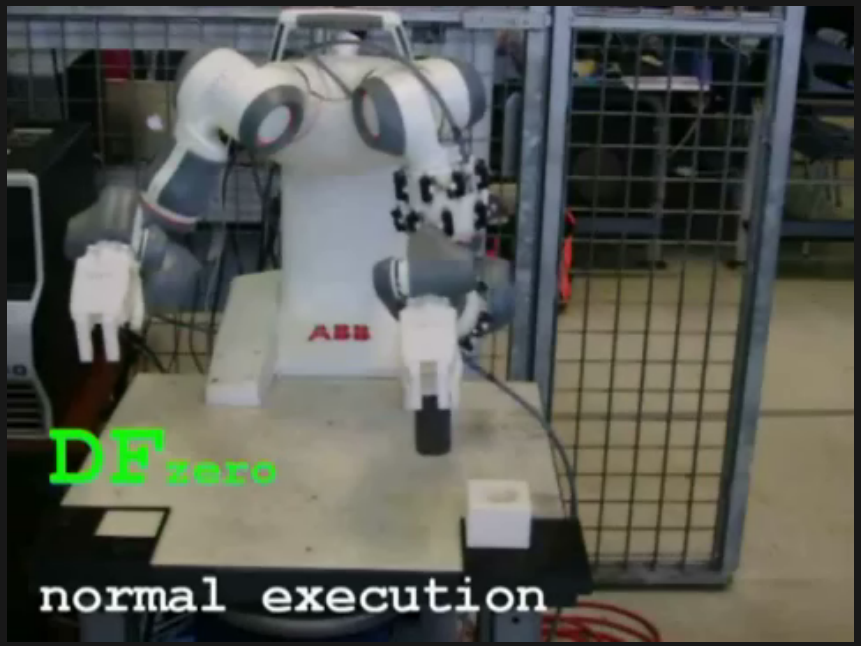
HRI is a (very) broad field

There are many different aspects of interaction between robots and people

- Interaction media
  - NLP
  - Communicative gestures
  - “Unconscious” gestures
  - Sounds
  - Lights
  - …
- Affection
  - Would we induce some affective effect in the user?
- Communicative goals
  - Information sharing
  - Request
  - Rapport
  - Persuasion

Some examples

Pet, sound-based, gestural and affective interface



Humanoidized arms, safe-perceived interaction

How can we benchmark HRI systems?

The basic principle

Experimental method → experiments to be designed to enable:

- Comparison
- Reproducibility / repeatability
- Justification / explanation

Competitions as benchmarking tools

Equating competitions to experiments is a debated issue

[Takayama, RSS Workshop on Good Experimental Methodology in Robotic, 2009]

However, there is an agreed-upon feeling that competitions can serve as benchmarking tools

In RoCKIn we take this stance and intend to use @Home competition to provide an operative approach to the benchmarking of, among others, some forms of human-robot interactions

RoCKIn@Home test bed (1)

"Granny Annie" lives in an apartment together with some pets. Granny Annie is suffering from typical problems of aging people.

RoCKIn@Home is looking into ways to support Granny Annie in mastering her life.

Tasks

- Getting to know my home
- Catering for Granny Annie's comfort
- Welcoming visitors

RoCKIn@Home test bed (2)

HRI in RoCKIn@Home

Three examples of human-robot interaction taken from the Rulebook of RoCKIn@Home

"Then Granny Annie lets the robot know that she wants to read, but cannot find her reading glasses at the bedside table. She asks the robot to find them for her"

"The robot opens the door, guides the Deli person to the kitchen, then guides him out again. The robot is supposed to always observe the stranger"

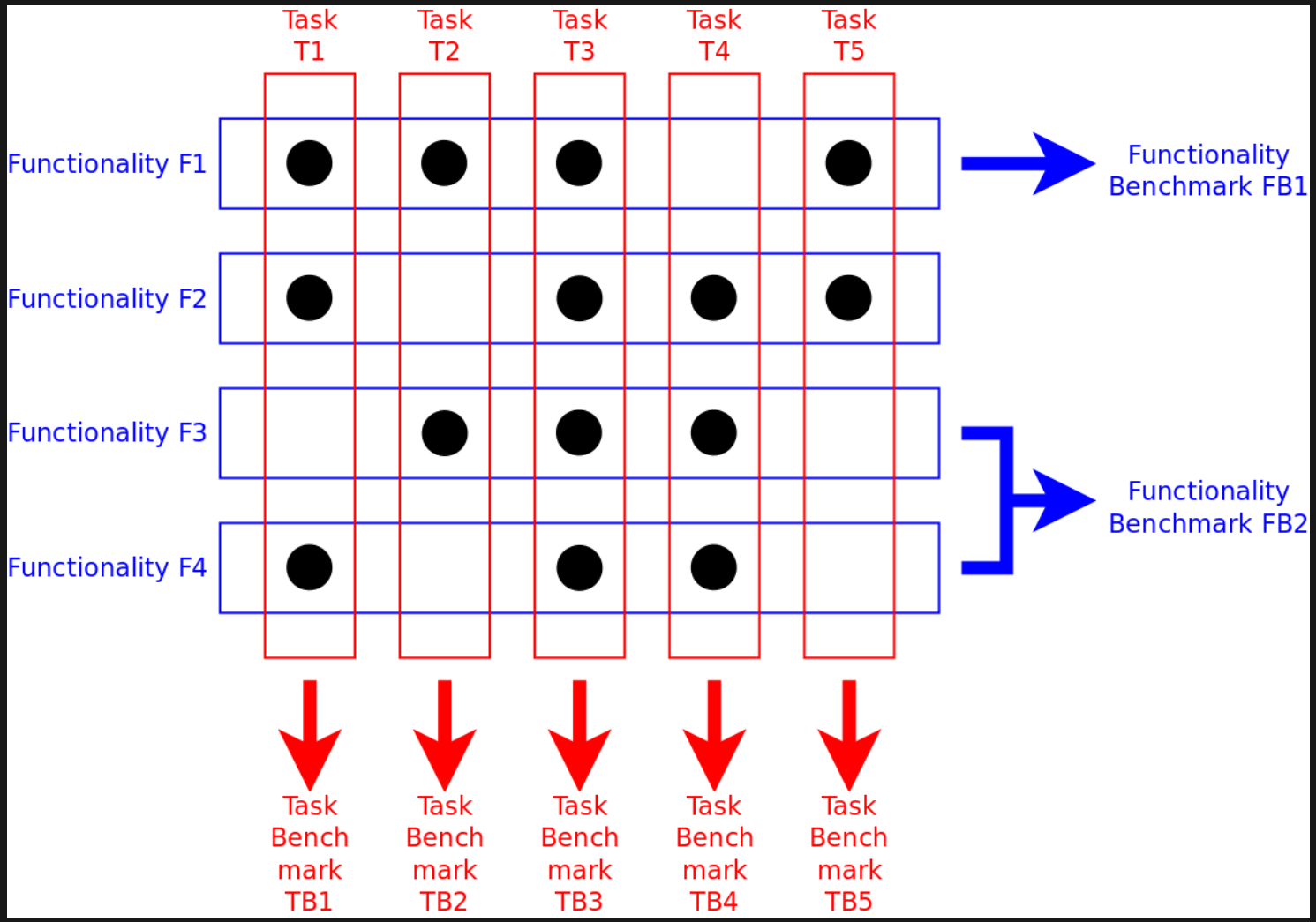
"Team members may "demonstrate" the apartment by guiding the robot through the apartment, pointing to objects and speaking aloud their names"

Benchmarking approach (1)

RoCKIn adopts a dual approach to assess robots performance in @Home:

- Task benchmarks
  - Getting to know my home
  - Catering for Granny Annie's comfort
  - Welcoming visitors
- Functionality benchmarks
  - Speech understanding

Benchmarking approach (2)



Getting to know my home
Task Benchmark

Achievements (yes/no):

- The robot detects the door with changed state
- The robot detects each piece of moved furniture
- The robot detects each changed object
- The robot correctly executes a command (issued either by voice or by following a team member) to move one the objects to a piece of furniture

There are also Penalized Behaviors (e.g., the robot requires multiple repetitions of human gesture/ speech) and Disqualifying Behaviors like when the robot damages the test bed

Catering for Granny Annie's comfort Task Benchmark

Achievements (yes/no):

- The robot enters the room where Granny Annie is waiting
- The robot understands Annie's command(s)
- The robot operates correctly the right device(s)
- The robot finds the right object(s)
- The robot brings to Annie the right object(s)

There are also Penalized Behaviors (e.g., the robot bumps into the furniture) and Disqualifying Behaviors like when the robot hits Annie or another person in the environment

Welcoming visitors
Task Benchmark

Achievements (yes/no):

- The robot reaches the door when the door bell is rung by Dr. Kimble and correctly identifies him/her
- The robot reaches the door when the door bell is rung by the Deli person and correctly identifies him/her
- The robot reaches the door when the door bell is rung by the Post person and correctly identifies him/her
- The robot reaches the door when the door bell is rung by an unknown person and correctly identifies the person as such
- The robot exhibits the expected behavior for interacting with the visitors

There are also Penalized and Disqualifying Behaviors

Speech understanding Functionality Benchmark

Measures used for scoring and ranking

1. The Word Error Rate on the transcription of the user utterances
2. The generated Command Frame Representations (CFRs) are evaluated (precision, recall, and F-measure) against the gold standard CFRs
    - Action classification (AcC)
    - Argument classification (AgC)
3. Time (if less than the maximum allowed for the benchmark)

The final score basically combines F-measures of AcC and the AgC (each contributing 50%)

Moving back to general case…

What would we like to benchmark?

- Goal achievement (answer: yes/no)
  - Does the robot understand Annie's command(s)?
  - Does the correctly identifies the Deli person?
- Performance (answer: a number)
  - Effectiveness (precision, recall, …)
  - Efficiency (time, resources used, …)
- Social and affective aspects (answer: a word)
  - Engagement
  - Trust
  - Compliance
  - Sympathy
  - Empathy

How can we benchmark?

- Goal achievement
  - Objective evaluation of goal achievement
- Performance
  - Performance objective measure
- Social and affective (possible at all?)
  - Panel of novice users, possibly trained to evaluate social and affective features
  - Possible use of qualitative scales (e.g., Lickert)
  - Possible use of evaluation grids
  - Possible consideration of both agents

[Leite et al., Int'l Journal of Social Robotics, 2013]

[Lee Koay et al., Int'l Journal of Social Robotics, 2014]

Thanks!