Seventh Framework Programme
Theme ICT-2009.2.1
Cognitive Systems and Robotics

HUMAVIPS
Humanoids with Auditory and Visual Abilities in Populated Spaces
Grant agreement number 247525

D7.4: Third project demonstrator (brief description of the prototype)

Date of preparation: March 2013

Contributors
Rodolphe Gelin, Aldebaran Robotics
Alexandre Mazel, Aldebaran Robotics
Jan Cech, INRIA
Soraya Arias, INRIA
Jordi Sanchez-Rieira, INRIA
Xavier Alameda-Pineda, INRIA
Antoine Deleforge, INRIA
Daniel Prusa, CTU
Vojtech Franc, CTU
Vasil Khalidov, IDIAP
Jean-Marc Odobez, IDIAP
Lars Schillingman, Bielefeld University

1. Introduction

To demonstrate the achievements of HUMAVIPS we have decided to continue with the scenario introduced in year 2 of the project, termed the “Vernissage". In this scenario (partially demonstrated during the second review meeting) NAO acts as a guide robot in a small art gallery taking care of a corner of the room where several paintings are exhibited. The robot is waiting in front of the paintings, being aware of the visitors in the room.

Based on the appropriateness given the current situation, the robot either takes the initiative to give explanations for some paintings or waits until being addressed by one of the visitors. This decision as well as keeping up the ongoing conversation requires good knowledge about the physical locations of the people in front of the robot as well as their audio-visual status: if a person both stares to the robot and emits a sound, there is clear evidence that this person is available for interaction. Besides being able to focus on a single person among a group, we have specifically chosen this scenario as it allows demonstrating the robot's ability to understand group constellations and react appropriately when multiple people are are facing the robot.

The targeted scenario is described in detail in the deliverable D1.5 “Scenario and detailed specification of M36 demonstrator”. In this present document we describe what has been implemented for the demonstration that is being presented during the final review of the project.

Formally (according to the Annex I of the HUMAVIPS project) D7.4 is a prototype to be delivered at M36. This prototype is available from the project's open-source dissemination website:

https://toolkit.cit-ec.uni-bielefeld.de/humavips/humanoids-auditory-and-visual-abilities-populated-spaces

The specific modules developed for the year 3 demonstrator are available at:

https://toolkit.cit-ec.uni-bielefeld.de/humavips/systems/humavips-y3-demonstrator

2. Description of the demonstrations

Figure 1 : Set-up of the demonstration

Figure 1 is a snapshot of NAO interactign with a group of visitors. It stands on a table to be seen by the visitors. The five pictures on the wall are the “paintings” that NAO should comment for the visitors. The robot is able to detect the group of person in front of it. It draws the attention of the visitors. When it has detected that one of the visitor is looking at him, NAO asks that visitor whether she/he wants some explanations about the paintings. Using the pictures themselves as landmarks, NAO is able to navigate towards any of the paintings and describe its content. During the visit, the robot maintains temporal information about the group of people (how many people?, where are their faces located? who is looking at the robot?, who is speaking?, etc.), and it regularly checks if a new person has joined the group.

In order to evaluate the different aspects of the interaction, we demonstrate the development of the project with three sub-scenarios. Each one of this sub-scenario focuses on a different aspect and shows specific functionalities associated with each module. For a real application of "Nao as a guide", the three behaviours which are presented in more detail below should be integrated within a unique scenario.

2.1 Communicative behaviour demonstration

Video of the demonstration:
https://www.dropbox.com/sh/99lb4rpvc2841b7/wWn70hR0rO#f:humavips-review-ald-demo-720p.mp4

In this demonstration, coordinated by Aldebaran, NAO proposes to start a new tour. It walks successively in front of each painting using the localization and navigation modules developed by CTU. Stopping in front of each painting, it gives an explanation about it, then walks to ne next one. When the robot stops, it has to look for the visitors in order to take a good position relatively to them and to the painting itself. When all the paintings have been described, NAO proposes a quiz to check if the visitors have memorized the information he just provided. Using the sound localization module, NAO detects who wants to answer its question. Its embedded automatic speech recognition (ASR) module interprets the answer given by the visitor such that NAO can comment the answer (see the Annex for the algorithm of the quiz).

This scenario demonstrates a real application of NAO in which the robot interacts with the environment (localization and navigation) and with people (speaker localization, speech recognition and dialog handling).

2.2 Attentive behaviour demonstration

Video of the demonstration:
https://www.dropbox.com/sh/99lb4rpvc2841b7/wWn70hR0rO#f:humavips-review-biu-demo-720p.mp4

In this demonstration, the robot explains the painting and goes from a painting to another, like in the previous one but the focus is put on the group management. The robot is able to check if someone is missing from the initial group. Then it can detect if the missing person comes back within the group or if a new person joins the group. In the demonstration, NAO adapts its behaviour accordingly to the evolution of the group.

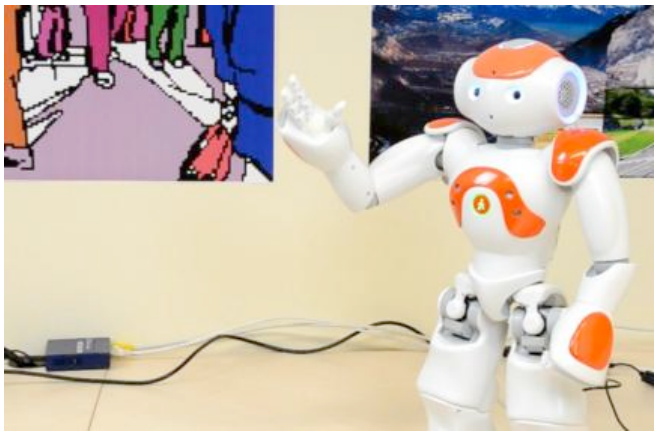

Figure 2 : NAO explains a painting

This demonstration is coordinated by BIU. It integrates the localization and navigation modules and the person and group manager modules.

2.3 Audio visual behaviour demonstration

Video showing the audio-visual fusion with head movements towards the speaking person:
http://www.youtube.com/watch?v=fMlz4IfYYGo

In this demonstration, NAO focuses on the speaking person by turning its head (azimuth and elevation) towards the face of the speaker. The audio-visual fusion method developed during the project uses auditory data from the robot's four microphones and visual data from the stereoscopic camera pair. When its attention is locked onto a speaker, he can recognize the person if his face is in the training data set. If the visitor claps his hands, NAO tells his/her name. To demonstrate the capabilities of the sound and face recognition module, we implemented a simple behaviour:

- if the visitor snaps his finger, NAO guesses his age by vision analysis.
- If the visitor clicks his tongue, NAO guesses his gender by vision analysis

This demonstration coordinated by INRIA relies on modules developed by all the partners. It enhances the scientific challenges tackled by the HUMAVIPS project in the domains of sound and visual processing.

3. Implemented modules

All the demonstrations are implemented on the RSB (robot service bus) middleware provided by BIU. This architecture allows the high-level modules to be executed on an external PC. Based on RSB, the modules developed by the partners could use the sensors on borad of Nao, particularly the stereo head developed by Aldebaran within this project. Detailed descriptions of the RSB modules can be found in the deliverables D2.2 and D2.3.

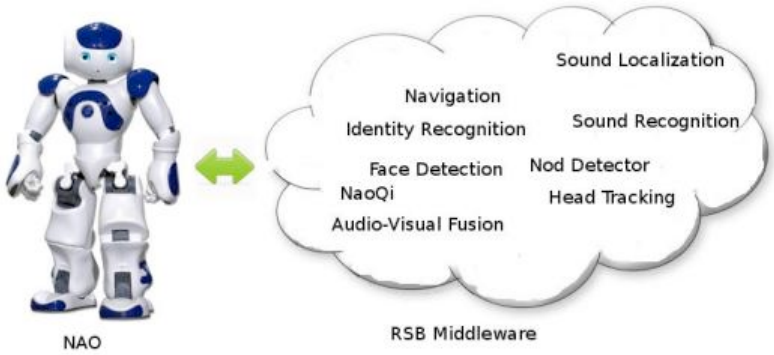

Figure 3 : Modules implemented for the demonstrations

In the following tables, we list the modules that were planned (described in deliverable 1.5). **Among these 28 modules developed by the partners, four of them could not be integrated in the final demonstrator**. The gesture recognition and detection module and the head gesture recognition

modules were tested using the project's data sets (RAVEL and VERNISSAGE) but their computational complexity did not allow a live RSB-based implementation. Similarly, the sound-source separation and localization module could not be implemented with the NAO robot head because the latter was not at all designed for source separation tasks (the robot's head-related transfer function, or HRTF, could not be properly modeled). Likewise, engagement tracking could not be implemented in live and interactive mode because of its computational complexity.

Visual modules	Available
Face detection	Yes
Identity recognition	Yes
People categorization	Yes
Simple gesture recognition and detection	No
Dense stereoscopic matching	Yes
Head gesture recognition	No
3D Localization and Tracking of Faces/Heads	Yes
Head pose estimation	Yes
Robot localization	Yes

Table 1 : visual functions

Auditory modules	Available
Sound recognition	Yes
Sound source separation and localization	No
Audio cue extraction	Yes
Association of Audio Cues with 3D Locations	Yes
Speech Recognition & Voice Activity Detection	Yes

Table 2: audio functions

Multi-sensor fusion and integration modules	Available
Person Manager	Yes
Group manager	Yes
Visual focus of attention estimation	Yes

Utterance	Yes
Addressee	Yes
Engagement tracking	No
Multi-party dialog management	Yes

Table 3 : Multi-sensor functions

Robot behaviour and coordination modules	Available
Scene manager	Yes
Task based behaviour abstraction	Yes
Navigation and motion control	Yes
High level behaviour coordination	Yes
Performance monitoring	Yes
Quiz	Yes
Explain painting	Yes

Table 4 : Robot behaviour functions

4. Conclusions

During the review meeting, the project demonstrators were run in an unexpected environment for which the methods were neither trained nor tested. The tested scenarios, while in an echoic environment, considered only a few people facing the robot. In addition, there were 10-12 people wandering and chatting around, thus producing a level of background noise that was much higher than expected. Nevertheless, the outcomes were quite good and the interaction performances (visual and auditory) between the robot and a person were not too much affected but these severe conditions.

Of course, sometimes the robot was not able to correctly understand what was going on, but humans have also difficulties to understand each other in such situations. On the positive side, it was demonstrated that the fusion of visual information with audio (which was the main scientific goal of the project) drastically improves the standard audio-only and speech-based human-robot interaction paradigm.

Some modules planned in D1.5 were not demonstrated (see Tables 1 to 4). The gesture recognition and the head position have not been demonstrated because they are computationally intensive and we did not have time to implement them properly onto the robot. They are very useful because it is

important for the robot to understand what the visitor designates by pointing the finger or the head. This will be developed after the project. The engagement tracking function, that is mainly based on head position, could not be evaluated either. This function could be useful but it has probably to be used carefully: the robot should not become a teacher that stops talking as soon as the visitor does not look at him anymore. At last, the sound source separation has not been implemented. This function would have been very spectacular (the robot could have been able to answer to two questions asked at the same time by two visitors) but it is very difficult to implement with the current robot head (see the comments above). If the robot does not understand a question because someone else speaks in the same time, it is acceptable that he asks to repeat the question. That is the reason why, from the application point of view, we could support to have this module not implemented within the final demonstrator.

Considering the ambitious objectives of the project described in the technical annex, one can say that they have almost been reached:

- A commercial humanoid robot is able to interact with a group of people, based on fused audio and visuas information, in a real physical world and in an unstructured environment.
- The robot is able to behave appropriately and to engage in a dialog with someone that is likely to be willing to interact with him.
- The robot can react to both verbal and non-verbal signals generated by humans who do not wear any special device and who are not in a specially equipped room.
- The robot can recognize humans (based on face, age, and gender) and to name them once they have been seen for a while.
- The robot technology that was developed relies on a memory-centred and cognitive architecture, which is being provided to the research community as an open-source software platform.

All these features point out the advantages of an audio-visual robot. The methodolgoy and technology developed by the project partners have been demonstrated in the "vernissage" scenario. This goes well beyond a standard proof of concept demonstrator: it is alrealy a real application for service robots. An interactive humanoid able to dialog with people and to have expressive gestures appears to be much more attractive, compared to existing museum guide robots.

The HUMAVIPS researchers are quite proud of their achievement. The integration of the demonstrations on NAO showed that the academic teams and the industrial partner can collaborate together to achieve an innovative application for a robot that is not just a laboratory prototype, but a commercial product.

Annex

1. Scenario of the demonstrator

The global description of the scenario is given on Figure 4. We have implemented this architecture on the Nao based on the development made by the partners.

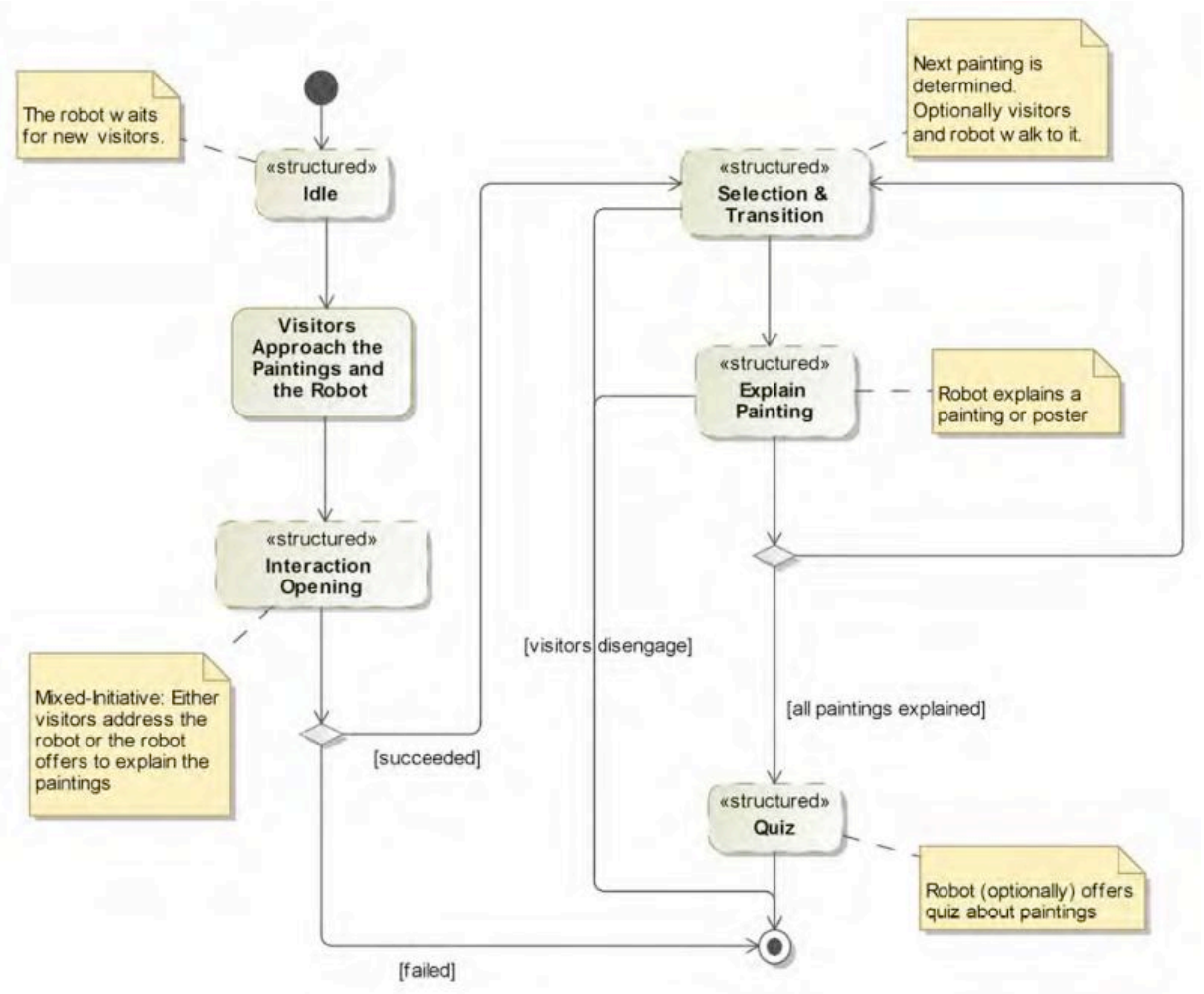

Figure 4 : Activity overview

The sub-scenarios “Idle”, “Interaction opening”, “Selection transitions”, “Explain painting” have been implemented accordingly to the description made in deliverable 5.1.

2. New version of the quiz

We have modified the description of the sub-scenario « Quiz » in which the robot asks questions about the painting. In the version described in deliverable 5.1, the robot used to select one visitor then ask him the question. In the version we propose, the robot ask the question to the whole group and he identifies who gave an answer from the group. It should make the interaction much more natural.

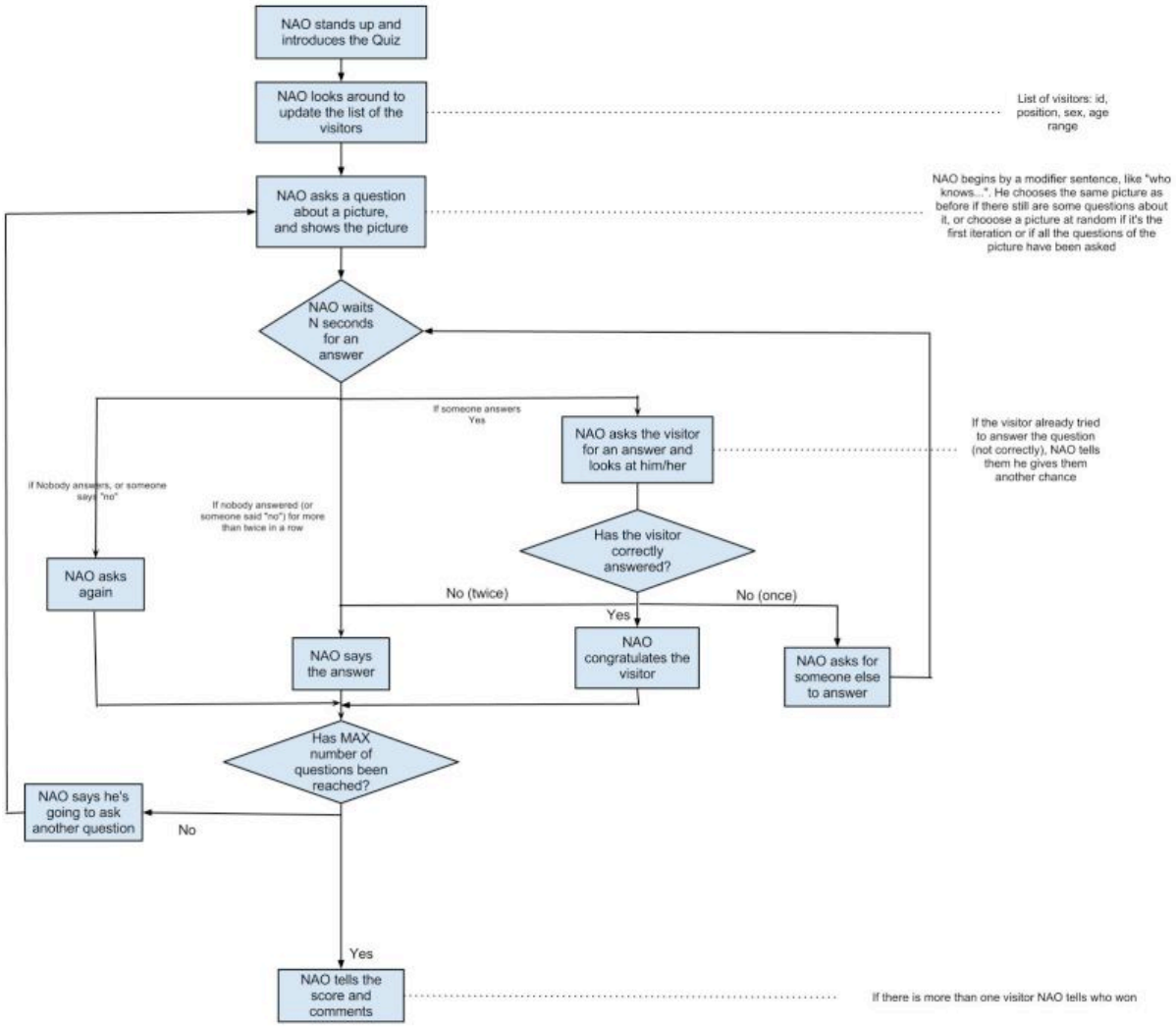

Figure 5 : New version of the Activity Quiz