Artificial Intelligence for Human-Robot Interaction
Papers from the AAAI 2015 Fall Symposium

Towards Gaze, Gesture and Speech-Based Human-Robot Interaction for Dementia Patients

Alexander Prange and **Takumi Toyama** and **Daniel Sonntag**
German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3
66123 Saarbrücken, Germany
firstname.lastname@dfki.de

Abstract

Gaze, gestures, and speech are important modalities in human-human interactions and hence important to human-robot interaction. We describe how to use human gaze and robot pointing gestures to disambiguate and extend a human-robot speech dialogue developed for aiding people suffering from dementia.

Motivation and Background

Dementia is a general term for a decline in mental ability severe enough to interfere with daily life. Alzheimer's is the most common type of dementia. Memory, thinking, language, understanding, and judgement are affected (Alzheimer's Association 2014). In our Kognit project we enter the mixed reality realm to help dementia patients.[1] In Prange and collaborators (2015) we introduced the use of NAO, a humanoid robot, as a companion to the dementia patient, in order to continuously monitor his or her activities and provide cognitive assistance in daily life situations.

Gaze behaviour plays a major role in human-human communication, hence effective human-robot interaction should include this modality, together with natural language and gestures. In this paper we extend our Kognit scenario and explore the use of embodiment to create effective multimodal human-robot interaction through the use of robot pointing gestures and human gaze to extend the human-robot speech dialogue. We refer to mild cognitive impairments (MCI).

Related Research

The gesture families presented in Jokinen et al. (2014) provide a background of robot gesture and speech capabilities, which we extend by adding pointing gestures and gaze orientation through eye tracking. Moyle et al. (2013) conducted a pilot study, in a residential care setting, showing that robot companions have a positive influence on the quality of life of people living with moderate to severe dementia.

A major challenge in human-robot interaction (HRI) is to achieve grounding in multimodal communication (human and robot use the same terms to denote physical objects). Mehlmann et al. (2014) explore a model of gaze for grounding in multimodal human-robot interaction which we take as a basis. Similarly, Broz and Lehmann (2015) propose a method for modelling conversational mutual gaze during turn-taking conversations in HRI and Yoshino et al. (2015) explore how to use robot gaze behaviour during human-robot conversations.

Gaze-based HRI

We use NAO in order to continuously monitor patients and provide cognitive assistance in patients' daily activities. Through microphones, cameras and sensors the robot companion is able to perceive its environment. Speech recognition and synthesis allow NAO to communicate with users through a natural speech dialogue. In our scenario the humanoid robot and the dementia patient share the same collaborative environment, a table they are sitting at. The user is wearing an eye tracker and the table is instrumented with an ARMarker sheet[2] to support recognition of gaze coordinates and thereby creating a shared perceptual basis. Using ARMarker for the table makes our approach scalable as the objects do not need markers themselves.

The user's eye tracker calculates the gaze coordinate in the scene camera image for each frame. Thus, we get the gaze position and the scene image for each frame. When the scene camera image captures one or more markers on the map, we recognise them and calculate the rotation and translation matrices of the markers. According to the rotation and translation matrices of individual markers, we can map the gaze coordinate in the scene image to a gaze coordinate in the ARMarker sheet. The gaze estimation method we use here is similar was presented by Vörös et al. (2014).

The NAO robot also has a scene camera on its head. Similar to the user's camera, we can also calculate NAO's gaze (head orientation) on the ARMarker sheet. We use the center of the scene camera as NAO's gaze point. If any markers are recognised in the scene camera, the humanoid robot can calculate its head orientation on the sheet. Since NAO's gaze and the user's gaze share the same ARMarker sheet, NAO can automatically rotate its head and point to the location the user is gazing at.

[1] http://kognit.dfki.de

[2] http://www.artag.net/

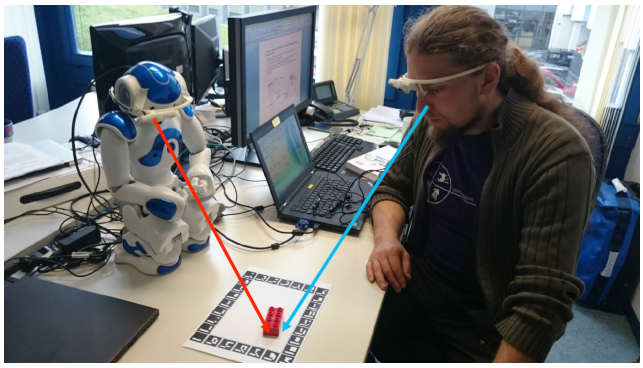

Figure 1: Gaze-based HRI in a shared interaction space

In order to provide efficient multimodal human-robot interaction we include human gaze (looking at an object) and robot gestures (pointing at an object) into our speech dialogue. In this way, referring expressions can be used to target a specific object among multiple objects on the table. To achieve grounding, the interaction include the objects in the instrumented area (see figure 1) and can be disambiguated automatically by resolving the eye gaze of the human together with the pointing gestures of the NAO robot. Thereby we create a shared interaction space between the robot and the human, allowing them to reference entities using a common set of terms.

Incorporation of Embodiment

During conversations, gaze behaviour plays versatile roles in collaborating with a shared workspace. Embodiment (here pointing gestures by the robot and human gaze feedback) plays a vital role in human-robot interaction and collaboration (Fang, Doering, and Chai 2015). In our dialogue the patient can refer to objects on the table by simply looking at them. Questions like "what is this?" can be asked while focusing a certain entity and the NAO robot then points to this object and describes it: "this is your pill box". When the dialogue manager recognises corresponding phrases, the robot requests XYZ gaze coordinates from the patient's eye tracker. It's head then moves towards the location and using the camera NAO looks for the object. Then the arm's joint motors are triggered, letting the robot's arm point at the recognised object.

More advanced scenarios include the use of the robot's background knowledge and the multimodal dialogue memory. Given the patient's healthcare plan, NAO can help the person take medications correctly either via speech-based reminders or direct pointing gestures on the pill box. The full Kognit storyboard can be seen in an accompanying video[3].

NAO: "Don't forget to take your pills." [*Patient points to yellow pills*]
PATIENT: "These?" [*NAO points to the blue pills instead.*]
NAO: "No, the blue ones over there."
PATIENT: "Do I take them with water?" [*NAO points to the water carafe.*]
NAO: "Yes, the water is right here. You should drink more, you haven't had water all day."

In addition to the speech dialogue with NAO, we use a stylus enabled smartphone to provide handwriting as another input modality during the dialogue (Prange et al. 2015). In cases where the user has to write something down, NAO points at the device and says "please write down your message using the pen", so that it is clear to the patient where to write. By providing this additional input modality we create a multimodal dialogue and circumvent the need to dictate free-form messages. If the device is located anywhere on the table (in the shared and marker-equipped interaction space) NAO can also react to "I can't find my phone" or "where is my phone?" by moving his head around and searching for the object inside the workspace. Once found NAO points to it "it's over there", otherwise he triggers the smartphone to play a notification sound, allowing the user to locate it in the room.

Architecture

Our current setup consists of four devices: the humanoid robot NAO made by Aldebaran Robotics[4], a Pupil Labs eye tracker[5] and a Samsung Galaxy Note 3 smartphone[6] with integrated stylus support. All devices are connected to a central server currently running on a desktop computer. Due to the distinct operating systems we decided to use XML-RPC, a remote procedure call protocol (utilising HTTP to transport XML encoded calls) that works cross-platform. Clients (devices) can communicate with each other using the server as a proxy (e.g., NAO requesting the user's gaze orientation). The server also keeps track of the dialogue history and the objects detected through object recognition.

The eye tracker and NAO are connected to a dialogue server on which the object recognition software and the dialogue platform are running. The object recognition is done by using the camera image of the user's eye tracker, cropping the local gaze region of the scene image (320x240) as described by Toyama et al. (2012). SIFT features are extracted from the local gaze region. The nearest feature vector in the database is retrieved for each SIFT feature and it casts a vote for each object ID. The object ID that has the majority of votes is returned as the recognition result. Note that we do not use the NAO's camera to recognise the object that the user gazes at.

Conclusions and Discussion

We described how to use human gaze and robot pointing gestures to disambiguate and extend a human-robot speech dialogue in an MCI dementia use case where especially memory about the location of objects and disambiguating the roles of these objects for daily life activities has a major role to play. While the object recognition delivers satisfying results, our first experiments show several issues in other areas. First of all the integrated camera of the NAO

[3] https://vimeo.com/132704158

[4] https://www.aldebaran.com/en/humanoid-robot/nao-robot
[5] https://pupil-labs.com/pupil/
[6] https://www.samsung.com/global/microsite/galaxynote3/

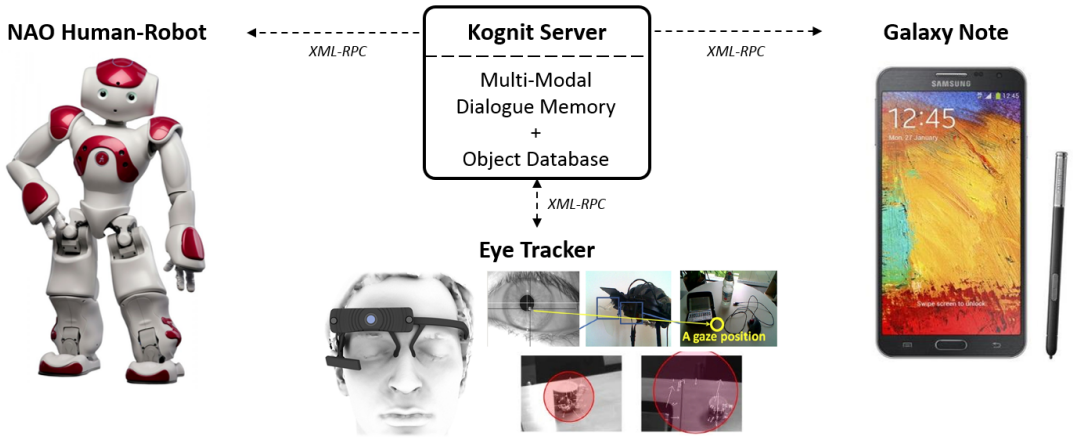

Figure 2: Kognit Architecture

robot delivers low quality images which lead to low recognition accuracy, so we had to attach an external camera to NAO's head. Second, although NAO has several joint motors the pointing and gaze range is very limited. Instead of pointing at the exact location of an object, we simulate this gesture with 6 fixed, predefined pointing gestures. Third, the eye tracker solution we use does not provide reliable gaze estimation. Mapping by single markers may be inaccurate, thus we average the gaze coordinates given by all captured markers. Since the perspective change of a marker in the scene image is large, it is hard to calculate correct rotation and translation matrices, leading to gaze coordinate estimation errors.

In the future we will work on solving these problems to create an effective multimodal gaze and gesture-based HRI scenario. An evaluation study will follow. More integrated scenarios have to be developed for dementia patients, e.g., to provide relief and more time for family members. Robots with appropriate gaze and robot pointing precision are key to help in addressing family issues related to caring (http://kognit.dfki.de).

References

Alzheimer's Association. 2014. 2014 Alzheimer's Disease Facts and Figures. http://www.alz.org/downloads/facts_figures_2014.pdf.

Broz, F., and Lehmann, H. 2015. A gaze controller for coordinating mutual gaze during conversational turn-taking in human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, 131–132. New York, NY, USA: ACM.

Fang, R.; Doering, M.; and Chai, J. Y. 2015. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, 271–278. New York, NY, USA: ACM.

Jokinen, K., and Wilcock, G. 2014. Multimodal open-domain conversations with the nao robot. In Mariani, J.; Rosset, S.; Garnier-Rizet, M.; and Devillers, L., eds., *Natural Interaction with Robots, Knowbots and Smartphones*. Springer New York. 213–224.

Mehlmann, G.; Häring, M.; Janowski, K.; Baur, T.; Gebhard, P.; and André, E. 2014. Exploring a model of gaze for grounding in multimodal hri. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, 247–254. New York, NY, USA: ACM.

Moyle, W.; Cooke, M.; Beattie, E.; Jones, C.; Klein, B.; Cook, G.; and Gray, C. 2013. Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial. *J Gerontol Nurs* 39(5):46–53.

Prange, A.; Sandrala, I. P.; Weber, M.; and Sonntag, D. 2015. Robot companions and smartpens for improved social communication of dementia patients. In *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, IUI Companion '15, 65–68. New York, NY, USA: ACM.

Toyama, T.; Kieninger, T.; Shafait, F.; and Dengel, A. 2012. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '12, 91–98. New York, NY, USA: ACM.

Vörös, G.; Verő, A.; Pintér, B.; Miksztai-Réthey, B.; Toyama, T.; Lőrincz, A.; and Sonntag, D. 2014. Towards a smart wearable tool to enable people with sspi to communicate by sentence fragments. In *Pervasive Computing Paradigms for Mental Health*, volume 100. Springer International Publishing. 90–99.

Yoshino, T.; Takase, Y.; and Nakano, Y. I. 2015. Controlling robot's gaze according to participation roles and dominance in multiparty conversations. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, 127–128. New York, NY, USA: ACM.