ICT-2009.7.1
KSERA Project
2010-248085

Deliverable D3.1

Human Robot Interaction

18 October 2010
Public Document

The KSERA project (http://www.ksera-project.eu) has received funding from the European Commission under the 7th Framework Programme (FP7) for Research and Technological Development under grant agreement n°2010-248085.

Project acronym:	KSERA
Project full title:	Knowledgeable SErvice Robots for Aging
Work package:	3
Document number:	D3.1
Document title:	Human Robot Interaction
Version:	9
Delivery date:	30 September 2010 (month 8)
Actual publication date:	18 October 2010
Dissemination level:	Public
Nature:	Report

Editor(s) / lead beneficiary:	Raymond Cuijpers (TU/e)
Authors(s):	David van der Pol (TU/e), Jim Juola (TU/e), Lydia Meesters (TU/e) Cornelius Weber (UH), Alex Yan (UH), Stefan Wermter (UH)
Reviewer(s):	Georg Edelmayer (TUW) Antonella Frisiello (ISMB)

Contents

Glossary of terms used

Acronym	Definition
HHI	Human Human Interaction
HRI	Human Robot Interaction
Naoqi	Client-server software architecture for programming Nao
Nao	Humanoid robot manufactured by Aldebaran Robotics
Paro	Therapeutic robotic baby seal
COPD	Chronic Obstructive Pulmonary Disease
Asimo	Humanoid robot developed by Honda
Kismet	Robotic head developed at MIT
Aldebaran Robotics	Manufacturer of the Nao robot
REC	Region of Eye Contact

Executive summary

This deliverable describes research advancing the state of the art to model social robot behaviour. The aim is to model the human-robot interaction (HRI) based on human-human interaction (HHI). The focus is on eye-contact and gaze behaviour during conversation. Human social communication patterns, such as engaging and disengaging in eye contact, head and gaze movements to follow a conversation attentively and gaze aversion caused by distracters in the environment are modelled and implemented on the Nao robot. The Viola and Jones algorithm and the built-in face tracking modules in Naoqi were investigated to model eye contact behaviour of the robot. Neural networks were used to improve and advance current state of the art methods for detecting a person's head pose. This enables the robot to judge whether a person is making eye contact or whether (s)he is attending to something else. Gaze aversion behaviour is based on detecting motion energy of objects in the robot's visual field. These algorithms were implemented on the Nao and the first implementation has been successfully demonstrated to the public. In addition, experiments were conducted to measure the tolerance in head orientation of the Nao for which a user still experiences eye contact. The first results showed that the tolerance region for which eye contact is experienced from the Nao robot is approximately similar to that experienced when interacting with another person.

Purpose of this deliverable

In this deliverable (D3.1) we report our first results on human robot interaction. The work reported here relates to task T3.1 Human-Robot communication & interface design (TUE, UH).

The research described in this report contributes to the usefulness and acceptability of robots in a home environment. The main focus of the research has been to model, implement and test robot bahaviour to facilitate joint attention.

Suggested readers

This document is recommended to all KSERA partners. The research is of interest to the scientific community as well as industry developing socially assistive robots. Any research or development in which the robot should cooperate with a human being can benefit form the research presented in this report.

Relationship to other documents

This document serves as input to task T5.1 and is part of MS4 which reports the intermediate results of WP1 through WP5. The part on “Head Pose Estimation for Real-Time Low-Resolution Video” has been published as a conference proceeding of the 28th European Conference on Cognitive Ergonomics (ECCE 2010) in Delft, The Netherlands, 25-27 August, 2010 (see also dissemination activities in the Progress Report 2010 and in D6.3).

1 Introduction

In this chapter we present background on social robotics and gaze tracking in particular. In Chapter 2 we describe the interaction model between a user and the Nao; i.e., feedback from the user and the different roles of Nao. Implementation of joint attention including attentive eye-tracking, attention monitoring and gaze aversion is described in Chapter 3. Chapter 4 presents the first results of the behavioural experiments on defining the eye-contact region for the Nao as observed by a human being. We conclude with Chapter 5, giving a brief outlook on the future work of work package 3.

1.1 Social Robots

Until recently, robotics has been mainly applied in industry. Examples of in-home use are toy robots (e.g., AIBO), domestic robots such as vacuum cleaners (e.g., Roomba®) and assistive robots as envisioned in KSERA. These in-home robots demand robot behaviour attuned to the users and their expectations, communications, and even feelings. Compared to normal electronic devices, people expect more from a robot.

1.1.1 Why social robots?

Currently work in robotics is expanding from industrial robots to robots that are employed in the living environment. The toy and domestic market as well as many projects focusing on assistive robots accelerate these developments. The introduction of robots into our homes calls for a proper interaction model between the robot and the user. Instead of the robot just running a pre-programmed behaviour without the possibility of interruption or other influence on the robot's behaviour, users should be able to control the robot in a natural, easy-to-use way. One way would be to use an interface, like a keyboard or touch-screen, but a more intuitive way would be to allow the user to interact with the robotic platform in a way that the user is already accustomed to; i.e., interacting with the robot as one would interact with other humans. The interaction would then consist of non-verbal and verbal behaviour. The robot should be able to participate in communication cycles while adhering to the norms and rules of Human-Human Interaction (HHI). Thus, the robot should be able to communicate socially.

Dautenhahn (2007) argues for the importance of social intelligence in robots. Current research predominantly focuses on sensory-based intelligence. Sensory based intelligence allows for basic intelligence like swarm behaviour, for example, when individual elements react to small changes in light intensity, etc. This type of swarm behaviour is associated more with the level of bacteria and insects. According to Dautenhahn (2007), social intelligence might be the key to taking robots to the next level of intelligence. The adaptation to social complexity has presumably caused the increase of overall intelligence in primate evolution (Dunbar, 2003). The increase in social skills resulted in the increase in volume of the neocortical part of the brain. Beyond the increased capabilities on a social level, the neocortex allows for reasoning that is more sophisticated. This analogy has prompted Dautenhahn (2007) to opt for the development of socially intelligent robots.

The human-robot interaction (HRI) within the KSERA project is inspired by human-human interaction (HHI). The goal is to equip the robot with some social intelligence so that it is capable of a more natural communication with the user. For KSERA a robot should adhere to the rules and cues that are natural from the perspective of the user instead of the user having to adapt to the capabilities of the robot.

1.1.2 Anthropomorphism and social behaviour

A good way to increase the acceptability of a robot for older people is to move it away from the technical group of devices and more towards the group of social entities. Interestingly, people tend to anthropomorphise complex objects (i.e., attribute animal or human-like qualities to them) and apply social models in order to predict their behaviour (Breazeal, 2003). The appearance of the robot has an effect on the degree of anthropomorphism. Nao, for instance, is a humanoid robot and is shaped similarly to a human. Users will therefore be more prone to attribute human-like qualities to it than to wheel-based robots. People are more willing to accept robots as social entities when they anthropomorphise them. As a result, the adherence to social norms by the anthropomorphised

robot is expected. Thus, the more human-like a robot appears, the higher are the expectations of people interacting with it (Tapus, Mataric, & Scassellati, 2007). It is therefore important to attempt to make these human-like robots behave socially, like people do. When robots do not live up to their expected social behaviour, the view that people have of them changes, pushing them back into the technical device group. For instance, Dautenhahn (2007) studied the involvement of robots with autistic children. She found that simple reactive behaviours of the robot are enough to let the children assign social intelligence to it.

It is therefore important that the KSERA robot; i.e. Nao, reacts appropriately to what the user does. Initially, the user is likely to anthropomorphise Nao because of its human-like appearance. The challenge is in enabling users to maintain that view throughout lengthy interactions. In addition to more socially complex behaviour, simple reactive behaviour, like following the user when he/she is passing by, promotes longer retention of this anthropomorphic view.

1.1.3 Effect of robots on older persons

Robots are known to have potentially positive effects on older persons. Wada, Shibata, Saito and Tanie (2002) showed that the interaction of an older person with their PARO robot increased their vigor. They also found that PARO was able to stimulate the older person to communicate with others.

According to Tapus et al. (2007) the amount of empathy that is communicated to the user is important. The robot should therefore appear to understand the emotions of the user and react to them in an appropriate way. Non-verbal and verbal reactions that are given to the user during, for example, exercises (and during other interactions as well) should always be attuned to the user's perspective. In the KSERA project, Nao might, for example, detect that the user has difficulties completing the exercise and react appropriately.

1.2 Head gestures

In order to enable social interaction the robot should be able to detect social interaction cues, and respond to them appropriately. The robot should also be able to use these social cues and gestures in such a way that the user understands them in the specific context in which they occur.

Robots are not able to use many non-verbal facial expressions that humans use. Breazeal (2003), Breazeal, Edsinger, Fitzpatrick and Scassellati (2001), and Breazeal (2002) modelled the robot behaviour of a more expressive robot, Kismet, which can direct its gaze using head orientation as well as eye-rotation, blink its eyes, move its head while keeping mutual gaze, raise and lower its brows and move its lips. The most important non-verbal communication is eye-contact and will be explored in this report.

Breazeal (2003) with Kismet and Mutlu, Hodgins, & Forlizzi (2006) with ASIMO have devised communicative models to improve social cues during HRI. Kismet uses behaviour from the subject as an input to regulate its own behaviour. Kismet has a few ways of passing turns to its social partner, one of them by making eye-contact with the participant (i.e., by looking him/her directly in the eyes). These turn-taking cues make the interaction more fluent, and as a result, the user is more engaged in the interaction. Thus, while Kismet is talking, it predominantly looks away from the participant. Mutlu et al., in their study using ASIMO, did not employ turn-taking. They tested the effect of mutual gaze (i.e., gaze directed at the face of the interaction partner) on the recollection of a story told by ASIMO. The authors concluded that mutual gaze increases the level of engagement in an interaction. Thus implementing the turn-taking cues that Kismet uses, could decrease the amount of engagement in the interaction, according Mutlu et al. (2006). In short, there seems to be a paradox. According to Breazeal (2003), using functional limited mutual gaze to make the conversation more fluent adds to the engagement, whereas according to Mutlu et al. (2006), having as much mutual gaze as possible results in the highest level of attention and therefore the highest degree of engagement in the interaction. Clearly there is some optimal level of mutual gaze to be determined in different communicative interactions.

1.2.1 Gaze

In human-human interaction, gaze direction serves as an important non-verbal communication cue. It has been shown that the head orientation of a robot influences the perception of gaze direction (Staudte & Crocker, 2009). Not only are users more engaged in the interaction when robots have their gaze fixed on the user (Sidner, Kidd, Lee, & Lesh, 2004), but the engagement is also influenced by conversational cues like joint attention (Kozima & Yano, 2001), turn-taking cues and head gestures. Such interactions require the robot to be able to detect the gaze direction of the user.

Mutual gaze also influences the attitudes that people have towards their interaction partner. Mason, Tatkow and Macrae (2005) showed a dependence of likeability and attractiveness on mutual gaze. Kleinke, Meeker, and Fong (1974) showed that people having continuous mutual gaze during a conversation are rated more favourably by observers, compared to people having no mutual gaze during a conversation. There is an important difference between the way women and men rate an interaction with direct gaze. Ellsworth and Ross (1975) let participants rate an interaction they had. Men, when looked at directly for the entire conversation, had a negative feeling about the conversation and their partners, whereas women had a more positive feeling about the conversation and their partners. Judgments of social skill are influenced by the usage of mutual gaze; more mutual gaze results in the perception of better social skill (Kuhlenschmidt & Conger, 1988; Kleinke, 1986). Mutual gaze also results in better information recall of the conversation (Fullwood & Doherty-Sneddon, 2006).

Gaze has a regulating function for turn-taking during conversation. It is important to know when it is your turn to talk, or when the other is finished talking. People tend to end their turn by looking at the other person (Kendon, 1967). At the same time, their interlocutor is often looking away while taking their turn. In effect, the person ending his or her turn is offering the floor to the other person by looking at the partner. Looking away has a self-serving function for the turn taker. By looking away, s/he tries to minimize the visual cognitive load otherwise induced by looking at the other person. Kendon (1967) found that for approximately 80% of the time a person has finished his turn s/he will gaze at the interlocutor. The other person when taking his or her turn will look away from the turn-giver approximately 80% of the time. Although the data varied strongly across individuals, Kendon (1967) also found that most people gaze away most of the time when talking, whereas listeners predominantly gaze directly at the other. The reason people tend to gaze away while talking has, as for taking turns, a cognitive reason. When speech is less fluent it is more common to look away to minimize the visual cognitive load and use all resources for finding the right words and structure. For social interaction mutual gaze is therefore important.

1.2.2 Nao as an interlocutor

Although Nao's physical appearance resembles that of a human (see Figure 1), its visual sensors differ from the human visual system. First, its main camera has a limited resolution and, second, the camera's field of view is rather narrow when compared to that of a human. For natural communication, these properties limit the capabilities of Nao. As we have already seen in HHI, people tend to look away during a conversation. Continuous staring elicits an uncomfortable feeling in the other person.

Figure 1. A person interacting with Nao

Nao should also alternate its gaze between gazing at the user's face and in a different direction. Contrary to humans, Nao will likely lose the view of the user while looking away due to its narrow field of view. As a result this complicates the implementation of these behaviours in Nao.

Wollaston (1824) noted that the perception of gaze direction depends on both the perception of head rotation and eye orientation. Figure 2 shows two faces both having the same eye drawings but a different head drawing surrounding them (i.e., everything except the eyes). If you look at the picture you can see the right person looking straight at you, simulating mutual gaze, the left picture is gazing to a point next to you. Thus, the eye-orientation together with the head rotation influences the gaze perception.

Figure 2. Wollaston effect.

Consequently, if Nao were to perceive the user's correct gaze direction it should determine both the head rotation and the eye orientation. The problem is that in order to determine the eye orientation it would need to determine the location of the pupil within the sclera. Small changes, changes smaller than a millimetre, in the pupil location have a large effect on the perceived gaze direction. For example, experiments have been done in close proximity to the camera (Breazeal et al., 2001) or with special eye-tracking equipment that cannot be attached to Nao (Yoshikawa, Shinozawa, Ishiguro, Hagita, & Miyamoto, 2006). With current technology it is not possible to detect eye-contact reliably unless the robot is close to the person.

Although looking at HHI is a good way to start the development of social robots, HHI is not entirely similar to HRI; people react differently to robots than they do to other humans. Probably, the most well-known observation is the uncanny valley (Mori, 1970) showing that humanoid robots may look like humans but when they do not behave like humans the experience becomes somewhat eerie. This is shown in the right panel of Figure 3 where the familiarity a person experiences is plotted as a function of the human likeness of moving (dashed line) and still (solid line) robots. Both curves clearly show a valley just before robots are close to being indiscriminable from humans.

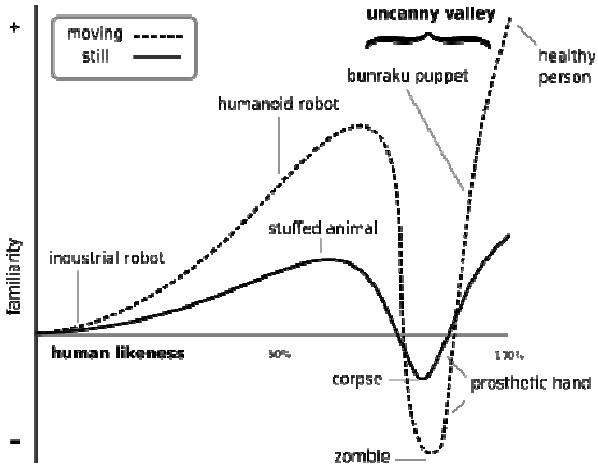

Figure 3 Geminoid of Prof. Hiroshi Ishiguro (left) and the uncanny valley in the curve relating familiarity to human likeness (right).

2 Interaction between Nao and a person

2.1 Autonomy and roles of Nao

In KSERA the Nao robot takes up different roles as outlined in the scenarios and use cases in D1.1. In accordance with that, Nao sometimes must take initiative itself and sometimes only respond to external events. The general aim here is that Nao acts as a fully autonomous agent and that external events like those stemming from ubiquitous sensing remain invisible to the person interacting with Nao. This leads to the following taxonomy of Nao's behaviours.

Idle operation: There are no external events from the KSERA intelligent server (see D2.1) nor does the person initiate an interaction with Nao. In this situation Nao autonomously decides what to do. For instance, Nao can go to sleep to conserve energy, go to its home location for recharging, monitor the person with its own sensors or initiate an interaction with a person. In the latter case the interaction may be simply to be responsive and alert or it may be more involved as in scenario 5: socializing and entertainment.

Nao is addressed directly by a person: If Nao is addressed directly (for instance by calling its name or patting its head) it should respond verbally and non-verbally by establishing eye contact with the person who called it. Clearly, the initiative is initially with the person interacting with Nao, but it may regain initiative when it decides to play its role as socializer and entertainer.

Scheduled events as outlined in scenario 1 (Healthy through indoor exercise), include reminding the person to perform a blood oxygen measurement or to carry out a pre-established set of fitness exercises. In these situations Nao is directed by the KSERA intelligent server to perform certain tasks, and Nao tries to obey as best it can. Although, strictly speaking Nao's actions are triggered by the KSERA intelligent server, this is invisible to the person interacting with Nao and it will seem that Nao acts of its own accord. Nao clearly has the initiative here.

Ubiquitous sensing may trigger events that require Nao to interact with the person. This situation is most similar to the scheduled events outlined above. For example, Nao may need to assist with disease self-management (scenario 2: Disease self-management), provide information about the air conditions (scenario 3: A safe environment) or anticipate on user needs (scenario 6: Smart home & navigation).

Medical emergency: In case of a medical emergency the KSERA system needs to alert the medical authorities (for example, Maccabi's call centre). Nao's role is limited to mainly reassure the person in need that help is on its way and/or to establish a direct video or audio connection with the outside world. In this situation Nao will play a passive role in the sense that it waits for commands to be given by the person in need or the medical authorities. The only autonomy that remains will be in the low-level behaviours like establishing eye contact and avoiding obstacles.

2.2 Feedback from Nao

Human-robot interaction is achieved, if the robot not only perceives information coming from a person, but also provides appropriate feedback to the human. A robot can follow a person and thereby always be in a position to communicate with him or her. The Nao has several effectors; in addition, in KSERA, it is augmented by a LED projector ("beamer"), which will be used to transmit messages to the user.

Thus far, an initial exploration has been made using the built-in communication facilities provided by the Nao robot in the context of face tracking. As soon as a face is detected by Nao, it can react by doing one of the following actions:

- Blinking eyes.
  An animation in the eyes will be launched using function *randomEyes*. The LED light diodes in the Nao eyes will emit a colorful light pattern.

- Bonjour using loudspeaker.
  Nao will say some kind words like: "Good morning!" using functions from the *ALTextToSpeech* module.

- Moving arms.
  Nao will move or wave its arms.

As said at the beginning, the feedback to the user is one of the more important subtasks of the human robot interaction. In addition to the Bonjour reaction, Nao can verbalise any information with the AlTextToSpeech module on demand. As we have demonstrated, it is possible to produce any feedback utterance by using sentences of common words. In its basic version, the module contains French and English vocabulary and voices. Since the module is based on the Acapela Mobility package, other languages like Italian and German are possible. In addition the module is able to queue several sentences received at any point of time to avoid losing single words or sentences. At this point these are mere possibilities, as they are not directly supported by Aldebaran Robotics.

To process spoken responses and utterances from the human, the Nao is capable of recognising words out of a fixed list of words with a variable confidence threshold. The microphones used on the NAO are not very high grade, which leads to a deterioration of the recognition capabilities with increasing noise and different human voices. By activating an implemented a High-Pass filter it is possible to cope with some of these problems and provide acceptable speech recognition under low-noise and clear-voice conditions. With this integrated speech recogniser, a simple dialogue like asking the human for a decision or simple statement is possible. For a more complex interaction like dialogue under normal (noisy) conditions a much more sophisticated system is necessary.

3 Implementing joint attention

We implemented three different behaviours that enable gaining, keeping and monitoring eye contact which are the essential first steps of joint attention: face tracking for gaining and maintaining eye contact, motion tracking for averting gaze, and head pose estimation for monitoring eye contact and joint attention in general.

3.1 Attentive eye-contact: Face tracking

Face tracking lets Nao detect the location of the user's face and direct its head towards the user. This gives Nao its most basic socially intelligent behaviour. It allows Nao to direct its speech to a specific person and, by doing this, keeps him/her engaged in the interaction. As discussed before, simply continuously looking at the user might elicit an uncomfortable feeling.

Two methods for face tracking were explored. The first method is based on the state-of-the-art face tracking method of Viola and Jones (2001). The second method is based on the face tracking algorithm integrated in Naoqi.

In the first method, we first gather the grey-scaled image from Nao. Currently the frame rate is limited by the connection speed. To solve this low-bandwidth problem we use a wired connection (instead of wireless) to connect to Nao and only one image channel (luminance) is transferred. The image is then searched for a face using Viola & Jones' (2001) method, which is available in the OpenCV library as the cvHaarDetectObjects() function. We use the python wrapper that is available for OpenCV. OpenCV is also provided with a file containing a cascade of classifiers that can be readily used for face detection. Viola & Jones (2001) employ an Adaboost classifier with Haar wavelet features. Although the Adaboost classifier makes the method usable within certain computational margins, it remains a computationally intensive method. To enable a faster implementation we limit the size of the window that is searched for faces. A face that is detected lies within a returned rectangle (Figure 4, white rectangle). If the previous image did not return a face, the next window within the image that will be searched for a face is 1.5 times as large as the rectangle containing the face (Figure 4, black rectangle). If this small window does not contain a face, the window size is increase for the next image that is retrieved (Figure 4, blue rectangle). If again no face is detected the window within the retrieved image will be even larger (Figure 4, red rectangle). This process is iterated until the window is bigger than the retrieved image. This speeds up the process significantly. Faces will generally be detected ten times a second (depending on the system). If there is no face detected earlier, and thus the entire image is scanned, the frame-rate is twice as low.

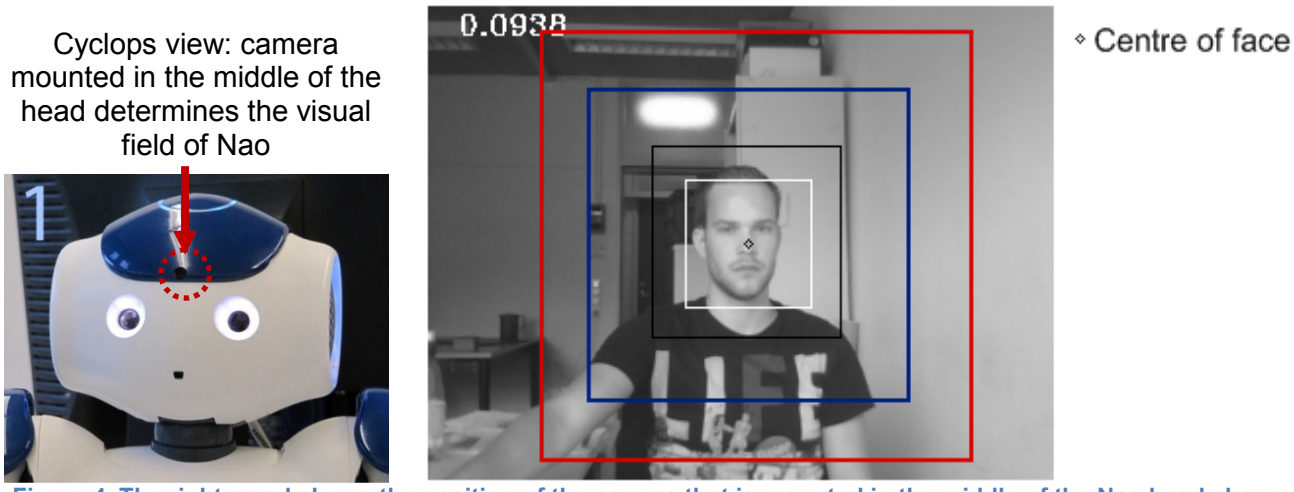

Figure 4. The right panel shows the position of the camera that is mounted in the middle of the Nao head above the eyes. The left panel shows the detection of a face in an image captured by the Nao camera.

The centre of the face (see Figure 4) is the point Nao should be pointing its head at. The goal is to get the centre of the detected face in the centre of the image. This assumes that the user will also perceive the robot to be looking at the point in space represented by the centre of the image.

In the second method, we use the face tracking algorithm that is integrated in Naoqi. The data transmission between the robot and the computer is avoided, which leads to a higher running speed. Firstly we use the face detection module *ALFaceDetection* in Naoqi.

The *ALFaceDetection* module is a vision module for the Nao robot to detect faces in front of its camera. It works well when the user looks towards the camera. Based on the detection results we use a closed-loop control circle for approaching the user; the work flow is shown in the following figure.

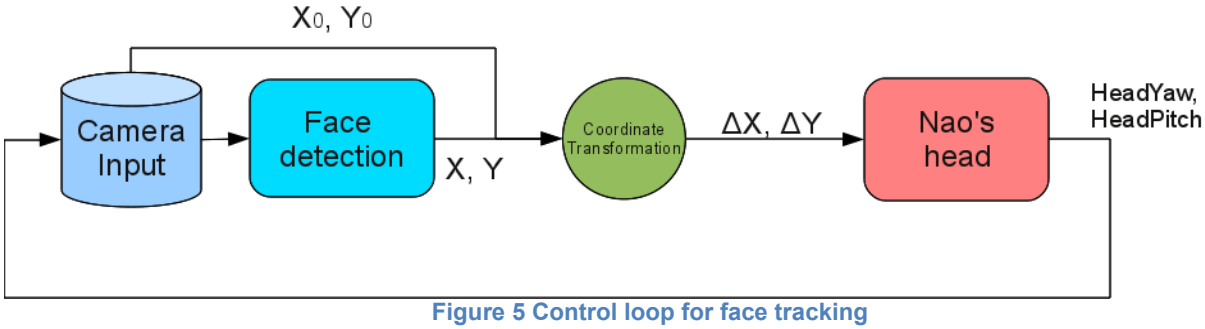

Figure 5 Control loop for face tracking

The centre point of the detected face is labelled here with a blue cross in the image and given in an X/Y coordinate system. The relative position to the centre point of Nao's camera will be transformed , and, according to this distance, Nao will turn its head (actor "HeadYaw" and "HeadPitch" in Naoqi) using the function *change_joint* to eliminate the deviation. The experimental results are shown in Figures 6 and 7. The red lines denote the centre of the image, and it can be seen that the robot now always turns its head to face the User.

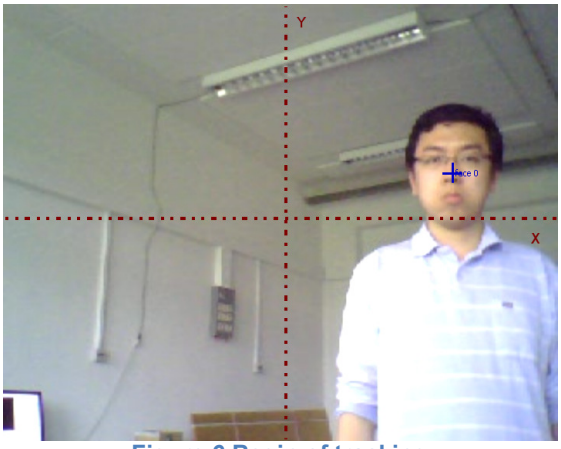

Figure 6 Begin of tracking

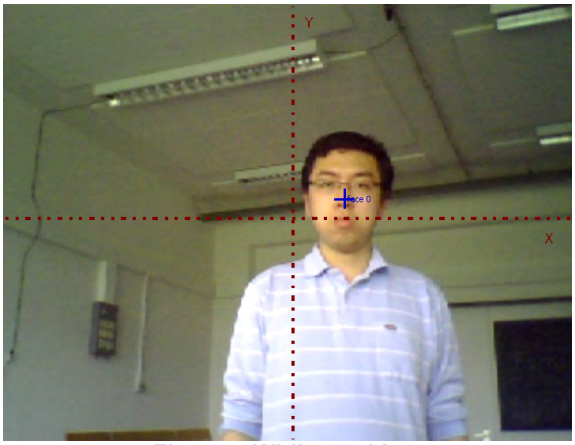

Figure 7 While tracking

One main drawback of this system is that the robot can only track one face even if more faces are detected in the image.

To move Nao's head we use the angleInterpolation() function. This function, provided by Aldebaran, moves the head of Nao to a certain position with respect to the current position. It does this movement within a certain time interval, which determines the speed of the movement. If the target position is relatively far from the current position, the speed of Nao's head movement is relatively fast. It enables Nao to track both slow-moving faces and fast-moving faces. During the movement the goal position is iteratively updated for each new frame. Thus, when Nao is getting closer to its

target position the head movement will slow down and end its movement in a smooth manner. If during a movement the face cannot be detected, it will continue to move in the direction of the previous movement, and the speed of the movement will reduce with each iteration. This way, when a face is moving too fast for Nao to follow, it will often still find the face.

Both methods of face tracking have their advantages and drawbacks. For more robust performance, both methods of face detection can be combined. In case of slightly deviating results (just a few pixels), the final face position can be determined as the average centre position of both procedures. In case of largely deviating estimates, further knowledge can be taken into account such as confidence values returned by the detection methods, or a plausibility check based on the previous positions of the detected face, assuming that face positions do not change by great distances between consecutive frames. Based on initial tests we will decide which approach to use in the first KSERA prototype. Once the face has been reliably detected in Nao's camera image, further analysis like gaze direction can be applied to the corresponding image region.

3.2 Gaze aversion behaviour: Motion tracking

Nao needs to be able to direct its gaze away from a person to other objects or to a random direction for a certain percentage of time during communication. To direct Nao's gaze away, we use a method for detecting movement energy from a pair of sequential images. This mimics the fact that, in human perception, motion easily attracts visual attention. Because the processing speed available on Nao is too limited to implement these functions in a real time fashion, it is implemented on the dedicated Nao server (see D2.1).

We use the motion tracking algorithm to enable Nao to maintain gaze at a moving object and to disengage from looking at a person. This way, we allow Nao to regulate the appropriate face-directed gaze with contextual face-averted gaze. For example, suppose Nao initially looks at the user while listening to him/her. After some time Nao should avert its gaze in order not to stare at a person. For instance, if Nao detects a moving car through the window, then Nao is able to use this to avert its gaze. It will find the car and track it along the road until it is out of view. Nao's gaze will then find the user's face again. As a consequence, Nao will stare at people but, at proper intervals, avert its gaze.

This behaviour is implemented in the following way: We start by determining the distribution of movement within the image collected from Nao. The previous image (t-1) is subtracted from the current image (t) (see Figure 8). If for a certain pixel image t-1 differs significantly from that same pixel in image t, an extreme value of 255 (white) represents that pixel in the difference image.

Figure 8 Difference image that extracts moving edges from an image sequence in this case of a hand moving from left to right.

These white parts in the difference image contain boundaries of the object that is moving. Boundaries are particularly good features of an object to track. Using the OpenCV function goodFeaturestoTrack() we will determine a few of these optimal feature points and save them in a list together with an initial weight value (1). These features are then searched for in the next image

(t+1) and their absolute movement is added to the weight value (see Figure 9). New feature points are added for each new frame. If the movement was only visible in one interval, the feature is removed from the list. If movements of features are large enough and keep occurring over time, their weights will eventually cross a threshold; enough movement is detected to ensure proper detection of a moving object. The median of all the feature points is used as the point that is eventually tracked by Nao. Tracking the object uses the same tracking function used to track faces.

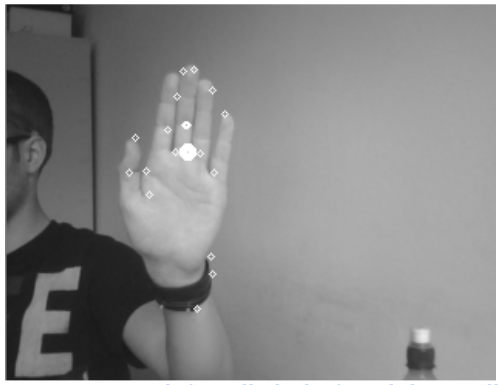

Figure 9 Image containing good features to track (small circles) and the median value of these features (larger circle).

3.3 Attention monitoring: Head Pose Estimation[1]

Eye contact involves two people looking at each other, or in our case, a robot and a person looking at each other. Face tracking allows Nao to look at a person but this person may not be looking back at Nao so that there is no eye contact. Thus, Nao should be able to infer the gaze direction of the person so that it can monitor whether there is eye contact and, if not, infer to what or where the person's visual attention is directed.

Presumably, the resolution of Nao's main camera is too limited to reliably detect the eye-orientation, but it can detect head orientation. Therefore we developed a head pose estimation method that measures pitch and yaw angles (see Figure 10). The roll angle can already be detected using the Viola and Jones (2001) method for face detection: If no face is detected there is either no face in the image, or the face has too much roll. In the latter case, the image may be counter-rotated and a new face search can be initiated. If a face is detected, the rotation of the image can be used as the approximate head roll. Currently, the face detection algorithm only works for a limited (but still considerable) range of yaw angles: the head pose cannot be estimated for a head that is seen *in profile*.

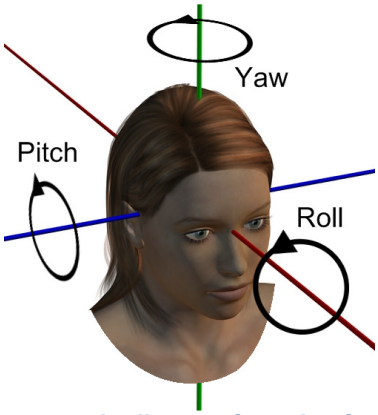

Figure 10 Pitch, yaw and roll axes of rotation for a human head.

We extended the work of Voit, Nickel and Stiefelhagen (2005). This was necessary because existing methods were not sufficiently precise and robust for our purposes. After first implementing

[1] This is an extended version of the ECCE 2010 conference paper (Pol, Cuijpers and Juola, 2010).

Voit et al.'s (2005) method for head pose estimation we found that their results are based on tests with stimuli from the same dataset with which they used to train their neural network. For heads of other people their method's performance was much worse. In addition, their dataset contains photos taken in an environment with constant lighting conditions. Again performance of their algorithm turned out to be unsatisfactory in varying lighting conditions. Other solutions using hybrid techniques to perform head pose estimation (Murphy-Chutorian & Trivedi, 2009) are capable of dealing with variable lighting conditions, but they do not work for low resolution images (or images in which the user is far from the camera). A commercial solution called *Face API*, performs well at close range, but when the resolution of the face decreases, it also stops functioning. Presumably, this is a consequence of the template matching algorithm looking for critical feature points within the face. Feature points can be used to create a 3D model of the face, but they are hard to detect in low-resolution images. Our approach uses the robustness for low resolution images from the Voit et al.'s (2005) method and improves the robustness for changes in lighting and for changes in head shape. Specifically, in the context of homecare it is important for the robot to detect the engagement of the user during interaction. With this in mind, we developed a fast way to estimate the head pose of the user. The highly variable lighting conditions within the home environment are taken into account as well.

3.3.1 Neural network solution

The neural network approach of Voit et al. (2005) is relatively fast and remains robust for low spatial resolutions. This allows detection of head orientation even when the person is far from the robot. To improve robustness for variation in lighting conditions, we used two different datasets to train the neural network: Yale's dataset with variable lighting directions (Georghiades, Belhumeur, & Kriegman, 2001), and the Face Pointing04 dataset (Gourier, Hall, & Crowley, 2004). For every image in each dataset, both the pitch and pan of the head were known.

We improved upon Voit et al.'s (2005) head pose estimation method in two ways: first, we improved the database of training examples and second, we used multiple neural networks that were trained with different initial parameters and different subsets of our improved database. Our database consisted of cut-out faces obtained by the Viola and Jones' (2001) object detection method. The cut-outs were first transformed to greyscale images, and then a Laplacian filter was applied to extract the edges. Unlike Voit et al. (2005) we did not use both the grayscale image and the edge detected image as inputs to the neural networks, because the edge detected images turned out to provide satisfactory results by themselves. For all database images we created training images of 30x90 pixels (see figure 11). We removed those images from the database for which the direction of illumination was from extreme angles (e.g., from below). Furthermore, extreme orientations for which the face was not entirely visible were removed. This resulted in a database of 5756 training images.

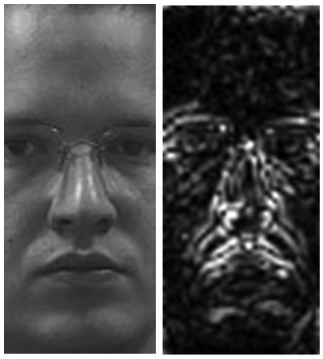

Figure 11 -out of the face obtained using the algorithm of Viola and Jones (2001) for object detection (left panel). Edges detected using a Laplacian operator that serves as the input to the neural network (right panel).

Instead of only one neural network we trained ten neural networks for estimating head pitch and ten more for estimating head yaw. In each set of ten neural networks three were trained using Yale's dataset (Georghiades, Belhumeur, & Kriegman, 2001), three using the Face Pointing04 dataset

(Gourier, Hall, & Crowley, 2004) and four using both datasets. We used two-layer, feed-forward neural networks which were trained using the Levenberg-Marquardt training method. All network weights were randomly initialised with values between 0 and 1.

3.3.2 Results

Our approach is able to give a good estimate of the head pose by averaging the responses of ten individually trained neural networks. The result is shown in figure 12 for estimating head pitch of a nodding movement. The top left panel shows the averaged network output as a function of time frame and the other panels show the individual network responses (labelled NN1 through NN10). Clearly, some networks performed worse than others. Five neural networks have a lot of high-frequency noise (NN 1, 2, 3, 4, 5, 7), one neural network (NN6) shows a reduced amplitude and the remaining neural networks perform rather well (NN8 through NN10). This pattern is typical for any head movement, but a priori it is impossible to predict which of the ten networks will perform best. Without external validation, as is the normal case, the best one can do is average across the individually trained neural networks, which is what we have done here.

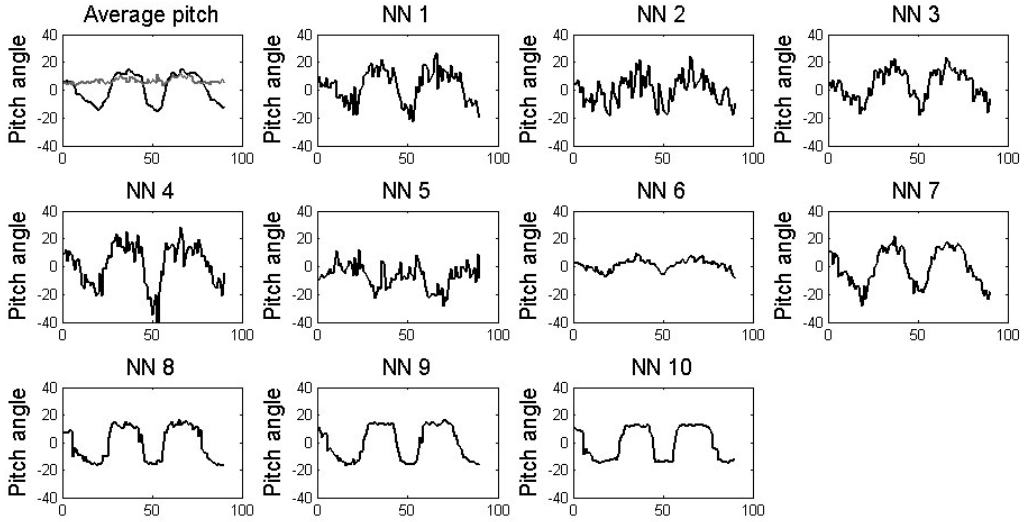

Figure 12 Each panel shows the detected pitch as a function of time frame. The top left panel shows the averaged pitch of a pool of 10 neural networks (black line) and the standard deviation is indicated by the gray line. The other panels show the outputs of the individual neural networks (labelled NN 1 to NN 10).

3.4 Conclusion

We have constructed a head pose estimation method that is optimised for low-resoultion images and that is robust over lighting changes. Compared to other methods our approach also performs better for faces that originally were not in the database. To do so we extended Voit et al.'s method (2005) by using an improved database of training images and by averaging the responses of a small set of differently initialised neural networks that were trained on different subsets of the improved database. The results confirm the improved robustness. With this method we are now capable of reliably estimating head poses of people from some distance away, which is precisely what we need for the KSERA project.

4 Behavioural experiment: Defining eye-contact for Nao

For Human-Human Interaction eye contact is well defined, but such data are lacking for Human-Robot Interaction because the facial features of robots typically strongly deviate from those of human faces. In this study we aim to measure the region perceived as eye-contact for the Nao robot.

4.1 Eye contact in Human-Human Interaction

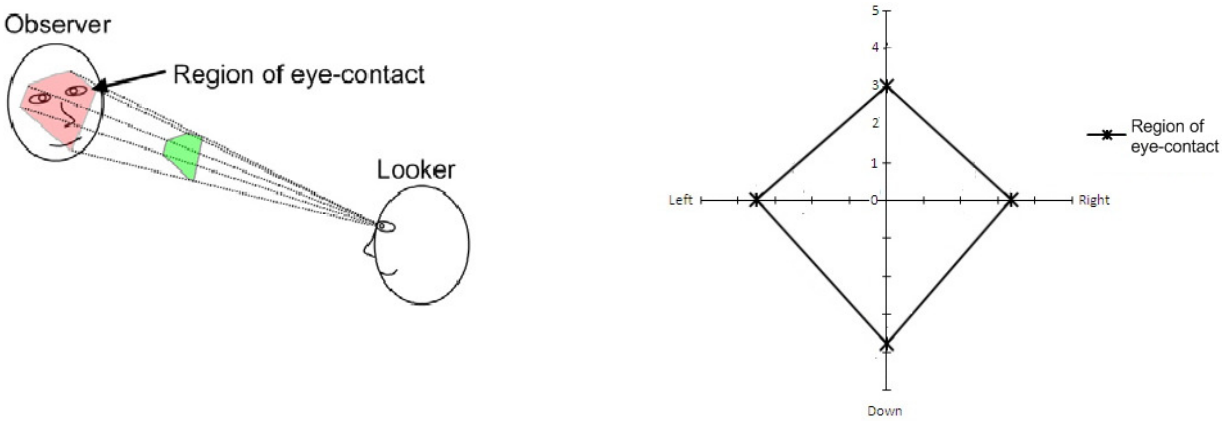

Figure 13 Region of Eye-Contact (REC). Left, the location of the REC on the face of the observer. Right, the size of the REC in degrees (van der Pol, 2009).

Since eye contact is so important in social interactions, many studies have reported on what gaze directions are perceived as eye-contact and what gaze directions are not. Observers of eye contact tolerate a deviation from a gaze directed at the nose bridge. This deviation defines a region, the region of eye contact (REC), which is slightly larger in the downward direction than in the other directions (see Figure 13). The REC defines a region on the face of the observer. If the looker is gazing within this region, the observer will perceive eye contact. The relative size of this region is expected to change with distance, although some deviations have been reported. When uncertain, perceivers are more prone to judge gaze as eye contact, resulting in an increase of size of the REC with distance. According to Chen (2002) the vertical asymmetry of the REC occurs because people reduce their eyelids separation when looking down and not when looking up. Closing the eyelids obscures the pupil position inducing additional uncertainty. As a result the REC will be larger in the downward direction.

Although many studies stress the importance of eye contact in HRI, studies that measure when eye contact is perceived with a robot are still lacking even though the facial features of robots, in particular the eyes, deviate considerably from human facial features. Here we measure the REC for the Nao robot. Nao's "face" features two dark holes that give the impression of eyes, and a small camera hole that is easily interpreted as its mouth (see left panel of Figure 4). Given these features and the symmetries of Nao's head Nao's head orientation is expected to be interpreted as Nao's gaze direction. Presumably, Nao's limited facial features and small head make judgements of gaze directions more difficult. Therefore we hypothesize that the REC for Nao will be larger than the REC for humans.

Nao is a *small* humanoid robot standing 56 cm tall. It is therefore likely that Nao will be looking up when making eye-contact. It is conceivable that this may have an effect on the shape and size of the REC. Therefore we also vary the height of Nao relative to the observer.

4.2 Method

4.2.1 Participants

Six participants took part in the study. Two of them were female and four of them male. Al participants were employees of the university and knew about the KSERA project, but they were not informed on the specific details of the experiment. The right eye was the dominant eye of all participants. Participants had normal or corrected-to-normal vision and were between 1.65 cm and 1.89 cm tall.

4.2.2 Apparatus

We used a 19" TFT screen to present the instructions to the participants. The humanoid robot Nao was used as the anthropomorphized looker. We recorded the values of the motor encoders, the completion time for each task, the size and position of the detected face as obtained by OpenCV's implementation of Viola & Jones' (2001) object detection algorithm, and the image from Nao's main camera at the moment that the "enter" key is pressed. Nao was either located on a small 40 cm high table (*standing* and *sitting* condition), or an 80 cm high table (*eye-height* condition). During the *sitting* and *eye-height* condition participants were seated in a chair that was adjusted to the same height for all participants. In every condition Nao was placed at 160 cm from the participant (see Figure 14).

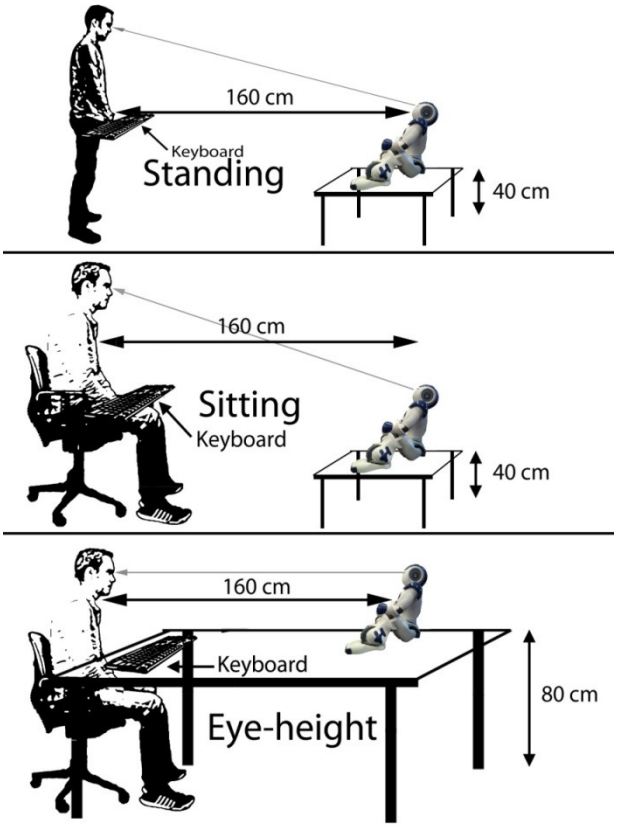

Figure 14 Experimental setup. Upper pane shows the setup for the *standing* condition. Nao is located on a small table while the participant is standing. Middle pane shows the setup for the *sitting* condition. Nao sits on a small table while the participant is sitting on a chair. The lower pane shows the setup for the *eye-height* condition. The participant is sitting while Nao is on the tabletop. Nao is located at eye-height.

4.2.3 Tasks

The participant's task was to adjust the head orientation of Nao until it appeared to look at the observer. There were four directions in which the participant had to adjust the head orientation (*up, down, left, right*). At the start of each session Nao told the participant in which direction to move Nao's head (see Figure 15). As soon as the participant perceived eye-contact (s)he pressed the "enter" key. Thus, the measured head orientation corresponds to the threshold of the Region of Eye-Contact (REC). There was also a task called *nose bridge,* in which participants were asked to direct

Nao's gaze towards their nose-bridge. This task started with Nao gazing beside the participant. The participant then moved Nao's gaze toward their nose-bridge adjusting both the yaw and pitch angles of Nao's head.

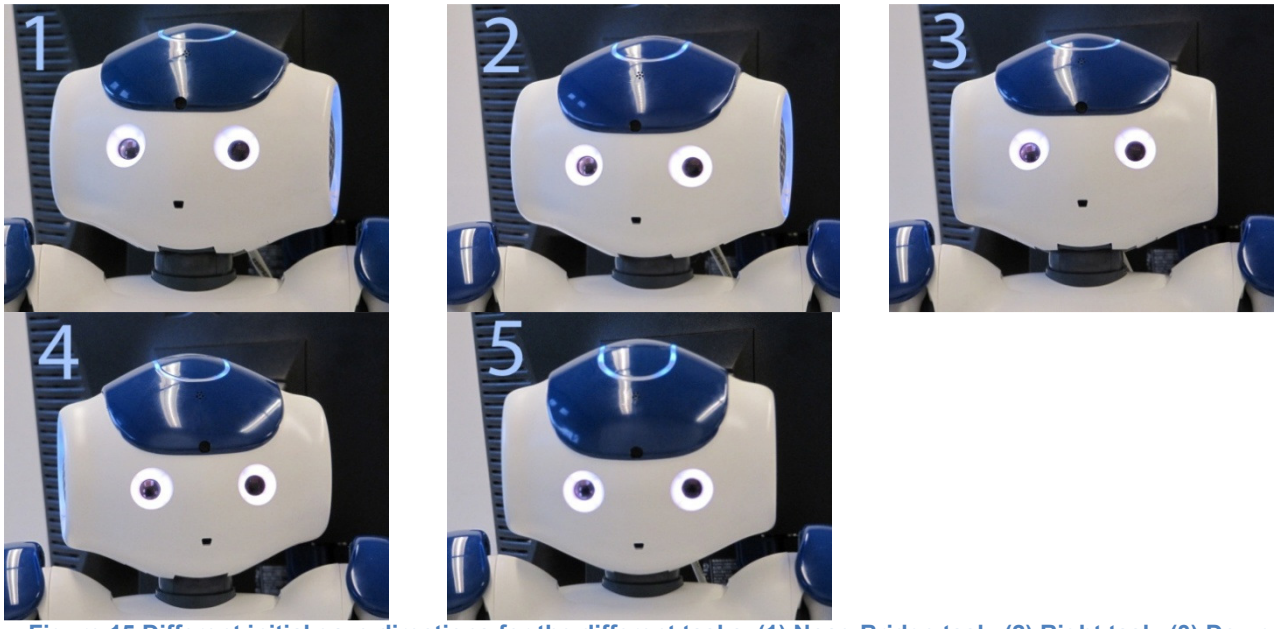

Figure 15 Different initial gaze directions for the different tasks. (1) Nose-Bridge task, (2) Right task, (3) Down task, (4) Left task, (5) Up task.

A within-subjects design was used for the experiment. There were three conditions: (1) a *standing* condition, (2) a *sitting* condition, and (3) a *eye height* condition (see Figure 14). For each condition, the REC was measured from four directions (see Figure 15) and each direction was repeated four times. The location of the nose-bridge was also measured from four directions, but each direction was repeated only once. This results in a total of 3 (conditions) x 4 (repetitions) x 4 (directions) + 3 (conditions) x 1 (repetition) x 4 (directions) = 60 trials.

4.2.4 Procedure

Every participant's body height was noted, and a test was done to determine the visual acuity and the dominant eye. Participants were asked to sit down on the chair and perform a few practice trials. The experimenter was be present during these practice trials to explain the different tasks. If the participant understood what to do for the different tasks, the actual experiment was started, and the participant was left alone. The experimenter could monitor the experiment's progress through Nao's camera.

4.3 Results

The REC gathered from the subjects was rather similar to what was found in previous experiments with human lookers. As for human lookers, subjects allowed for more deviation from looking at the nose-bridge, in the downward direction (see Figure 16). The centre of the REC differed from the measured nose bridge location: the pitch angle was significantly less, $t(71) = -4.85$, $p < .05$, but the yaw angle was not significantly different, $t(71) = .30$, $p = .77$.
The width of the REC at its widest point was M = 3.7 deg, SD = 1.3 deg and its vertical span was M = 5.0 deg, SD = 1.5 deg. The size of the REC did not differ significantly ($p>0.56$) from the ones obtained for human gaze experiments (5.14 degrees horizontally and 4.94 degrees vertically).

The effect of the conditions (standing, sitting or eye height) are shown in Figure 17. The height of the REC was larger than the width but the conditions standing sitting or eye height had no effect on the size of the REC ($F(2,71) > 0.2$, $p > 0.82$).

4.4 Discussion

4.4.1 Human gaze perception compared to Nao's gaze perception

The most important conclusion from the results of the experiment is that the perception of the gaze-direction and specifically, eye contact with Nao as a looker, is not that different compared to a human looker.

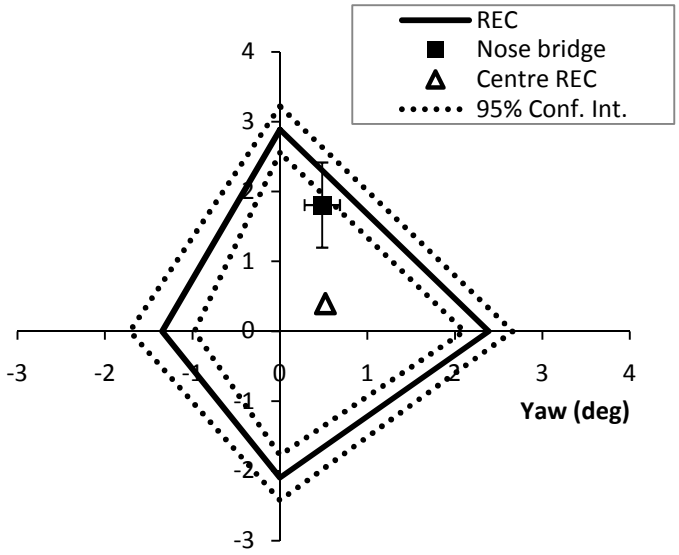

Figure 16 Region of eye contact (REC, solid line) from the participants' point of view. The center of the REC (triangle) does not coincide with the origin, but it is horizontally aligned with the gaze directed at the observer's nose bridge (square).

As for human lookers, perceivers allowed for more deviation when Nao was looking down than when it was looking up. Thus, contrary to what Chen (2002) proposed, this asymmetry is not caused by the increased uncertainty, due to the eyelids making eye-movements more salient in the upward direction, simply because Nao does not have eyelids. More likely, the asymmetry is caused by the socially common behavior that people show when looking at one another. The eyes, or nose bridge, are not centered in this region, and therefore are not the center of the REC (see Figure 16).

4.4.2 Calibrating Nao

There is a clear, significant difference in the horizontal axis between the origin of the graph, the middle pixel in Nao's video frame, and the measured centre of the REC or nose bridge. This deviation is due to the alignment of the camera of Nao and the subjective direction of Nao's head. After investigation, a significant difference was found between the two. However, this misalignment could not explain the location of the center of the REC.

The obtained REC constrains the parameters for Nao's eye contact behavior. For the experiment, we used the centre of the image coming from Nao as the presumed gaze direction. The nose of the participant would be in the centre of the image. Current results show that the point that should be in the centre of the image, which is the centre of the REC (see Figure 16), lies approximately 0.52 degrees (2,5 pixels in a 320x240 window) to the left and 0.39 degrees (1.9 pixels in a 320x240 window) above the centre of the image coming from Nao. In terms of eye contact, it is better to use the centre of the REC Nao's gaze direction because it divides the downward and upward direction in equal parts, making the chance of accidentally breaking eye contact as small as possible.
The REC enables the use of different head orientations all signaling eye contact. We can use this property to signal emotions using head orientation. For example, looking down is perceived as having a *sad* emotional state, where looking up is perceived as having an *"up"* or *alert* emotional state.

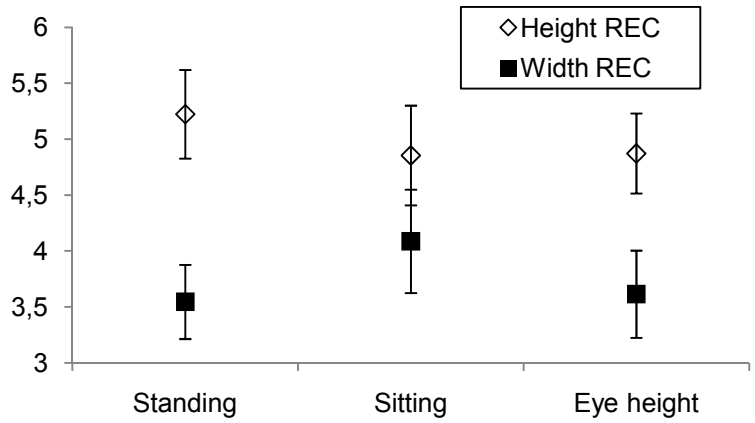

Figure 17 The vertical (diamonds) and horizontal (squares) dimension of the REC plotted for every pose of the participant. The whiskers represent the 95% confidence intervals.

4.5 Conclusion

The results from this study provide a good estimate of the tolerances for eye contact with the Nao. They show that these tolerances are similar whether the person is standing, sitting or at eye height with Nao. Our results also show that care should be taken to use the centre of Nao's camera images as looking straight ahead. Preferably, one should calibrate the straight ahead direction for each Nao robot in order to obtain the best results. The results from this study also tell us how to break eye contact. The obtained tolerances give us the precision requirements for Nao's gaze behaviours. For example, they allow us to signal emotion using head tilt without losing eye-contact with the observer.

5 Outlook

We have implemented several elementary behaviours on Nao that are considered crucial for any human-robot interaction. The next challenge will be to integrate and co-ordinate these behaviours with each other and with the navigation behaviours developed in WP2 for the first prototype.
Scientifically, much work must still be done to characterize what looking behaviour is actually experienced as natural and how it can be used for the more general notion of joint attention. This applies to both human-human and human-robot interaction.

In addition to the research reported in this deliverable the integration of a LED projector on the Nao is being explored. These results will be demonstrated and reported in the next prototypes (deliverables D3.2 in M18 and D3.3 in M30). The integrated LED projector is an extension to the current functionalities and affordances of the Nao. This approach keeps the advantages of a humanoid appearance such as affective behaviour and extents it by being able to present audio-visual information close to the user.

Interfaces like a traditional TV-set, a novel IPTV, a PC, a Laptop or a PDA are either static in their environment or need to be carried by the user. In both cases the user has to approach or carry the device to access the information. Then again mobile autonomous navigating devices like a robot can bring the information close to the user independent of his or her position. The user benefits from this because the information/communication link is presented not only when it is needed but also where it is needed. In an emergency situation the older person might be turned away from the traditional displays in the home and unable to access these. In such a situation the Nao brings the information to the user. Another advantage, for instance in the case of video communication, is that the experienced presence including a feeling of proximity, trust and naturalness with the displayed communication partner can be tuned by positioning the Nao at an appropriate distance to the user.
Task 3.2 will explore the technical feasibility of a LED beamer mounted on the Nao and also consider the usability, usefulness and acceptability of such a novel display solution. This includes a comparison of user preferences to traditional displays and the mobile autonomous information/communication delivery by the Nao. As a participative design is adopted in this project, user tests will be conducted to obtain a clear understanding in which situations the LED projector on the Nao is preferred, has an added value, or is the only communication/information link to the outside world in comparison to traditional displays in the user environment.

6 References

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2008). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71-81. doi: 10.1007/s12369-008-0001-3.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, *42*(3-4), 167-175. doi: 10.1016/S0921-8890(02)00373-1.

Breazeal, C., Edsinger, a., Fitzpatrick, P., & Scassellati, B. (2001). Active vision for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *31*(5), 443-453.

Chen, M. (2002). Leveraging the asymmetric sensitivity of eye contact for videoconference. *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*, 49. New York, USA: ACM Press. doi: 10.1145/503376.503386.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human-robot interaction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*(1480), 679-704.

Dunbar, R. (2003). THE SOCIAL BRAIN: Mind, Language, and Society in Evolutionary Perspective. *Annual Review of Anthropology*, *32*(1), 163-181. doi: 10.1146/annurev.anthro.32.061002.093158.

Ellsworth, P., & Ross, L. (1975). Intimacy in response to direct gaze. *Social Psychology*, 592-613.

Fullwood, C., and Doherty-Sneddon, G. (2006). Effect of gazing at the camera during a video link on recall. Applied Ergonomics, 37(2), 167-175. doi:10.1016/j.apergo.2005.05.003

Gamer, M., & Hecht, H. (2007). Are you looking at me? Measuring the cone of gaze. *Journal of experimental psychology. Human perception and performance*, *33*(3), 705-715.

Georghiades, a., Belhumeur, P., & Kriegman, D. (2001). From few to many: illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 643-660. doi: 10.1109/34.927464.

Gourier, N., Hall, D., & Crowley, J. (2004). Estimating Face Orientation from Robust Detection of Salient Facial Features. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*. Cambridge, UK.

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, *7*(2), 135-169. doi: 10.1075/is.7.2.04kap.

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta psychologica*, *26*(1), 22.

Kleinke, C., Meeker, F., & Fong, C. (1974). Effects of gaze, touch, and use of name on evaluation of. *Journal of Research in Personality*, *7*(4), 368–373.

Kleinke, C. (1986). Gaze and eye contact: A research review. *Psychological Bulletin*, *10*, 78-100.

Kozima, H., & Yano, H. (2001). A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, 47–52.

Kuhlenschmidt, S., & Conger, J. (1988). Behavioral components of social competence in females.

Mason M.F., Tatkow E.P., Macrae C.N. (2005). The look of love: Gaze shifts and person perception. Psychological Science, 16, 236–239.

Mori M. (1970). Bukimi no tani [the uncanny valley]. *Energy, 7,* 33–35. (In Japanese).

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: a survey. *IEEE transactions on pattern analysis and machine intelligence*, *31*(4), 607-26. doi: 10.1109/TPAMI.2008.106.

Mutlu, B., Hodgins, J., & Forlizzi, J. (2006). A storytelling robot: Modeling and evaluation of human-like gaze behavior.

Pol, D. van der (2009). The effect of slant on the perception of eye-contact. *Master Thesis*.

Pol, D. van der, Cuijpers, R.H., Juola, J.F. (2010). Head Pose Estimation For Real-Time Low-Resolution Video, ECCE 2010, 25-27 August 2010, Delft, The Netherlands.

Sidner, C. L., Kidd, C., Lee, C., & Lesh, N. (2004). Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces*, 78–84. ACM New York, NY, USA.

Staudte, M., & Crocker, M. W. (2009). Visual attention in spoken human-robot interaction. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction - HRI '09* (p. 77). New York, USA: ACM Press.

Tapus, A., Mataric, M., & Scassellati, B. (2007). The grand challenges in socially assistive robotics. *Robotics and Automation Magazine*, 1-7.

Viola, P., & Jones, M. (2001). Rapid Object Detection using a Boosted Cascade of Simple. In *Proc. IEEE CVPR 2001* (Vol. 01).

Voit, M., Nickel, K., & Stiefelhagen, R. (2005). *Neural Network-based Head Pose Estimation and Multi-view Fusion.*

Wada, J., Shibata, T., Saito, T., & Tanie, K. (2002). Analysis of factors that bring mental effects to elderly people in robot assisted activity. In *Proceedings of the 2002 IEEE/RSJ, Conference on Intelligent Robots and Systems*, 1152-1157. Lausanne, Switzerland.

Wollaston, W. (1824). On the apparent direction of eyes in a portrait. *Philosophical Transactions of the Royal Society of London*, *114*(1824), 247–256.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., & Miyamoto, T. (2006). Responsive robot gaze to interaction partner. In *Proceedings of robotics: Science and systems*.