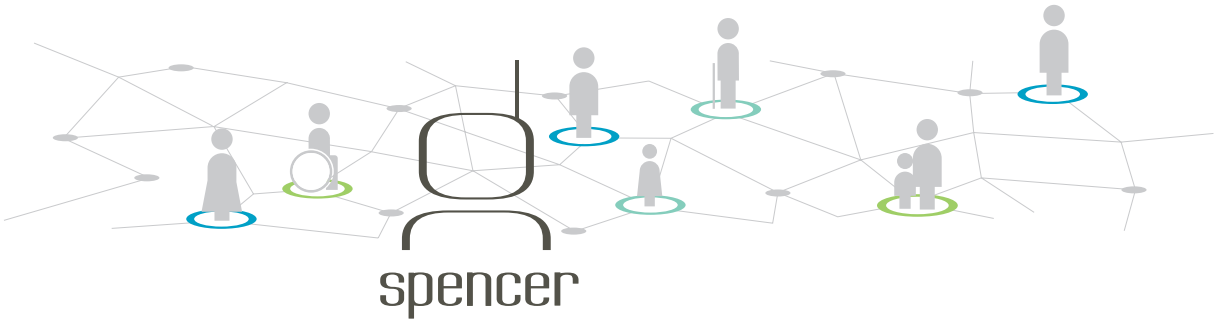


Grant agreement no: FP7-600877

SPENCER:

Social situation-aware perception and action for cognitive robots

Project start: April 1, 2013

Duration: 3 years

DELIVERABLE 6.5

Final evaluation report

Due date: month 36 (March 2016)

Lead contractor organization: UT

Dissemination Level: PUBLIC

Main Authors:

Michiel Joosse (UT)
Patrick Balmer (BLUE)
Rudolph Triebel (TUM)
Lucas Beyer (RWTH)
Stefan Breuers (RWTH)
Luigi Palmieri (ALU-FR)
Timm Linder (ALU-FR)
Tomasz Kucner (ORU)
Martin Magnusson (ORU)
Kai Oliver Arras (ALU-FR)

Version History:

0.1, initial version, mj, March 2016
1.0, initial version, mj, April 2016
1.1, added sections 2.2 and 2.4, pb, April 2016
1.2, changed structure, rt, April 2016
1.3, added sections 3.1-3.5, lb, lp, tl, tk, mm, May 2016
1.4, final tunings, koa, rt, May 2016

Contents

Figure 1: SPENCER robot at Schiphol Airport during final deployment in March 2016.

Abstract

This deliverable reports on the evaluation of the SPENCER platform at Schiphol Airport during the final integration week V in March 2016. We will report results based on the the evaluation criteria set forth in T6.4 (technical, scientific, end-user and user-experience measures). In summary, during the integration week the SPENCER robot did not show any technical failures, and the major system components such as people tracking or close-range perception performed better than the state of the art, compared on benchmark datasets. The robot travelled a total distance of $46.36km$ autonomously. In addition, the user studies conducted with 18 participants showed overall participants' appreciation for the robot and provided valuable insights to direct future research in domestic service robots.

1 Introduction

This report documents the final and iterative evaluation of the Spencer robot with regards to technical- and user experience tests. During the project lifetime, two technical integration and test events have been conducted at the test site, Schiphol Airport, in November 2015 (integration week IVb) and March 2016 (integration week V).

Two user tests with the Spencer platform have been conducted with users, collecting subjective assessment of the robot's capabilities. The first test was conducted with 29 small groups in a semi-public university environment, documented in Appendix A. The second series of user tests were conducted at Schiphol Airport during the integration week V in March 2016.

In this report we describe the Spencer platform (Section 2), technical tests (Section 3) and user experience tests at Schiphol (Section 4). Conclusions are presented in Section 5. The robot will be evaluated regarding several success measures as specified on page 6 of the Description of Work:

1. Technical measures (Section 2.4)
2. Scientific success measures (Section 3)
3. Use-cases specific impact- and subjective user experience measures (Section 4)

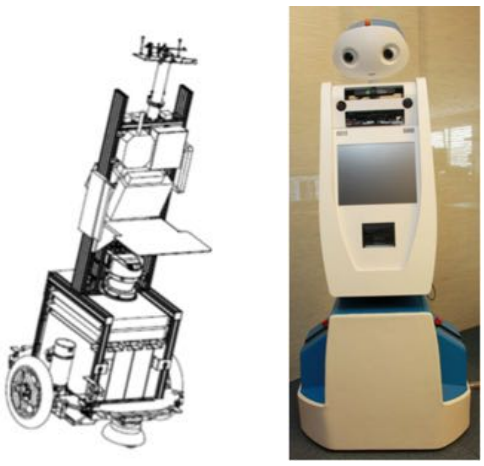


Figure 2: Spencer base platform without (left) and with (right) anthropomorphic shell

2 The SPENCER platform

The design of the SPENCER platform called for an aesthetically-appealing platform, equipped with hardware allowing the robot to autonomously navigation for an extended period of time. In this section we will summarize the technical specifications of the SPENCER platform (Section 2.1), detail the safety devices of SPENCER (Section 2.2), present an overview of the mission terrains (Section 2.3) and finally present the technical succes measures (Section 2.4).

2.1 Robot platform technical specifications

During a workshop session at the start of the project, documented in D1.1, specifications for the SPENCER platform have been defined. These specifications include among others:

1. Motion in crowded areas
2. Autonomy; operate at least 1 working day[1] without being recharged
3. Nice aesthetically designed, in line with KLM brand values
4. Maximum diameter to pass through standard doors, into elevators and through crowds.

Based on these specifications, a robot platform was developed (documented in D1.2, D1.3 and D1.4). The robot base platform and robot with anthropomorphic shell can be seen in Figure 2. The main technical specifications of the platform are as follows:

- Kinematics: symmetric differential drive with castor wheels
- Dimensions (h x l x w): 1926 x 810 x 800 mm.

[1] 8 hours

- Maximum velocity: 1.6 m/s [2]
- Acceleration[3]: 0.9 m/s^2
- Power autonomy: 8 hours
- Maximal step climbing: 10 mm.

2.2 Safety devices

The SPENCER platform is equipped with four safety devices:

1. **Wireless emergency stop button**. Emergency stop button connected wireless to the platform. When pressed, it opens it relay. Certified device.
2. **Emergency stop button (2x)**. Common industrial emergency stop buttons. When pressed, they open the safety loop.
3. **Bumper (2x)**. The bumpers are mounted on a spring with a switch. When pressed the electrical signal can trigger a relay.
4. **Additional safety relay**. Relay controlled by a PC via a USB port.

As can be seen in the Figure 3, all the safety devices are connected in series and if one triggers (i.e. open its relay) the so-called safety loop opens and the platform stops immediately. Moreover the doors housing the hardware are equipped of switches that will trigger the safety loop when opened.

During review meeting II, the reviewers asked for a safety audit to control that the initial measures were sufficient. During this audit it was shown that the software should also be able to trigger the safety loop, when detecting an obstacle. It was decided to add a safety relay (point 4) directly controlled by the software. The safety audit has been documented and reported in D6.6.

2.3 Mission terrains

The SPENCER robot operates indoors at Schiphol Airport, which features various environments posing challenges to the robot (both hardware- and software wise).

In the evaluation scenario, Spencer navigated through lounges and corridors (Figures 4-6), each with different surfaces and lightning conditions. Part of the floors are made out of black shiny material which is especially difficult for the lasers to detect.

Additional challenges which characterize the operating environment include conveyor belts and obstacles such as benches, information signs and artworks (Figure 5).

In the Schengen-Gate guidance scenario, Spencer awaits passengers behind the Schengen barrier, guides the passengers though Lounge 1, through a corridor to the B-piers, where the robot delivers the passengers at the correct gate, such as gate B18.

[2] Tested at 1.6 m/s, maximum operating speed 1.3 m/s for safety reasons

[3] While moving forward, not when turning

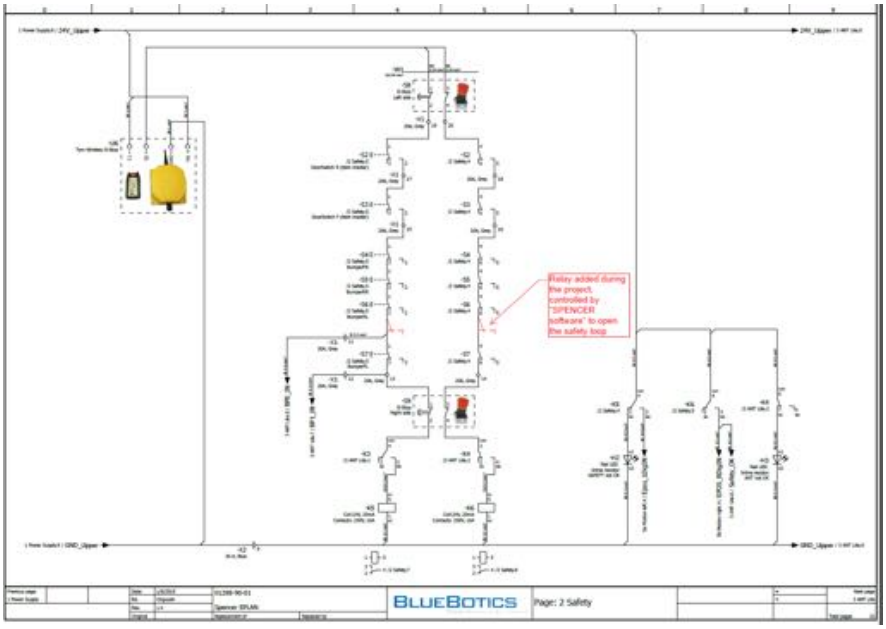


Figure 3: Schematic showing safety devices of Spencer. If one device is triggered the platform stops immediately.

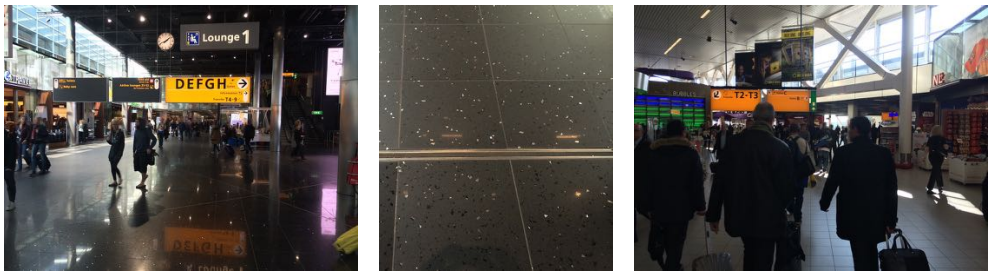


Figure 4: Lounge 1 can be characterized as a semi-crowded area with various intersection pedestrian flows. Additionally this area (partially) has a black shiny floor. Immediately following Lounge 1, the pedestrian flow splits towards piers B and C. Transfer T2-T3 are located in this area.

Figure 5: Both B- and C-piers contain gates on the left and right side, divided by two moving walkways. Due to the width of the paths these easily get congested.

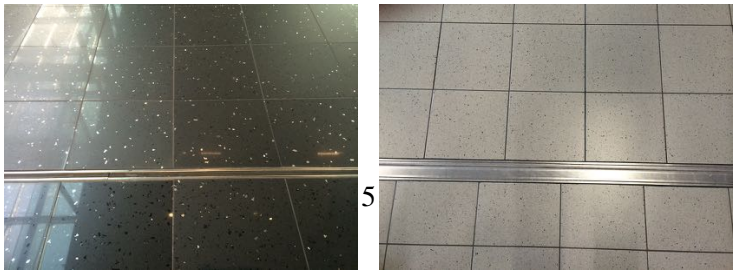


Figure 6: Various small thresholds can be found in the floors

Date	Problem/Solution		Email	On-site	Development	Improvement	Support
10-12.03.2015	Replace ANT lite	Support in Toulouse (Mathias)		x	x		
22.04.2015	Safety issue, improve braking distance	email from Kai	x		x		
28.04.2015	Lasers not synchronized	email from Timm	x			x	
28.04.2015	Laptop heating heating too much	email from Timm	x			x	
28.04.2015	Head pan joint dammaged (head moving freely)	email from Timm	x				x
03.06.2015	batteries deeply discharged	email from Harmish	x				x
04.06.2015	Head pan joint dammaged (head moving freely)	email from Rachid	x				x
05.06.2015	Support request to solve previous problems	email from Rachid	x				x
22-24.06.2015	Head: repair pan axis and add spring for tilt axis	Support in Toulouse (Gaëtan)		x		x	
22-24.06.2015	Lasers: synchronized them	Support in Toulouse (Gaëtan)		x		x	
22-24.06.2015	Safety: validate braking distance	Support in Toulouse (Gaëtan)		x		x	
22-24.06.2015	Batteries: reanimate batterie pack and install voltage meter	Support in Toulouse (Gaëtan)		x			x
08.07.2015	Support request during pre-integration week	email from Timm	x				x
10-12.08.2015	Laptop: add fans to plate above	Support in Toulouse (Gaëtan)		x		x	
10-12.08.2015	Safety loop: add USB relay	Support in Toulouse (Gaëtan)		x		x	
10-12.08.2015	ANT: connect directly to PC1	Support in Toulouse (Gaëtan)		x		x	
10-12.08.2015	Head: rotated by 180° as main driving direction will be backwards	Support in Toulouse (Gaëtan)		x		x	
31.08.2015	Supply CAD file for the 3D Velodyne support	email from Tomasz	x			x	
27.09.2015	Head position repetability bad	email from Harmish & Rachid	x				x
12-14.10.2015	Head: repair pan joint axis	Support in Toulouse (Gaëtan)		x			x
12-14.10.2015	Head: advice to decrease acceleration/deceleration to reduce effort on joint axis	Support in Toulouse (Gaëtan)		x			x
1-2.12.2015	Head: repair tilt axis (one screw missing)	Support in Schiphol (Gaëtan)		x			x
04.02.2016	Head: exchange head in Freiburg	Support in Freiburg		x			x
16.03.2016	Head: issue in the tilt axis of the head	Email from Timm	x				x
22.03.2016	Head: issue in the tilt axis of the head, controller problem	Email from Timm	x				x

Table 1: Overview of support requests to BLUE.

2.4 Robot platform success measures

BLUE had the task to evaluate the system based on technical measures. The evaluation sessions were short, as most of the time was taken for development and integration. Metrics measured during these evaluation sessions would not have been relevant. In order to get data as relevant as possible, long term metrics had to be defined.

It was decided to use *support requests* and *interventions* as metrics to evaluate the system on technical measures during the project. The reliability of the platform can be traced from the delivery of the platform to the final evaluation. It is also possible to analyze if the support requests were for support (e.g. broken part), for improvement (e.g. bug tracking) or for new development (e.g. improve braking distance).

Table 1 shows all the support requests chronologically. The date is indicated, the origin of the support request and the topics. The indicated support, improvement and development has been added.

2.4.1 Additional developments

Interventions and support requests for additional developments happened twice and only before the integration week III. The first development was the exchange of the ANT box in order to provide

SPENCER with the latest hardware. The second development was the improvement of the braking of the platform as first tests showed that the braking was weak and could therefore lead to safety issues.

2.4.2 Improvements

Between the integration weeks III and IV, a period of intense work on the platform, several support requests were sent to BLUE for improvements on the platform:

- Synchronize the laser scanners in order to avoid having interference between them.
- Laptop compartment too hot: the laptops were crashing after getting hot in the compartment. BLUE decided to add two fans to extract hot air, which solved the problem (Figure 7).
- Safety. Validate braking distance. Tests were performed in Toulouse, to be sure the braking distance of both platforms were adequate and sufficient to deploy the platform in Schiphol.
- Safety loop, add a relay: the risk analysis showed that the software should be able to trig the safety loop, which was solved whit adding a USB-relay controlled by the software and triggering the safety loop.
- Connect ANT directly to PC1: In order to decrease the latency in the navigation command, it was decided to connect the ANT box directly to one of the PC's instead connecting it through the switch.
- Tighten the head's degrees of freedom (DoF). A spring has been added to the tilt axis.
- Rotate head: the consortium decided to move the robot backward (in order to keep the display facing the passengers), the head had to be moved by 180 degree.
- Velodyne support: the consortium decided to add a Velodyne sensor in order to increase the perception (See D2.7, Section 2.1 for additional information). BLUE designed the support which has been printed with a 3D printer by ORU.

After the integration week IV, no more improvements were requested, mainly for the reason that the system had to be stabilized rather than improved.

2.4.3 Issues on the platform

BLUE also received support requests for issues on the platform between integration weeks III and IV:

- One battery pack was deeply discharged, BLUE had to wake up the battery.
- The head pan axis was damaged. The motor was moving freely several times and was repaired.

The support requests after integration week IV were only related to the head. BLUE decided in January 2016 to exchange the head, in order to use the new head on the platform in Schiphol during

the final deployment. During integration week V at Schiphol, the tilt axis had to also be repaired once, and the tilt axis controller was in error once as well.

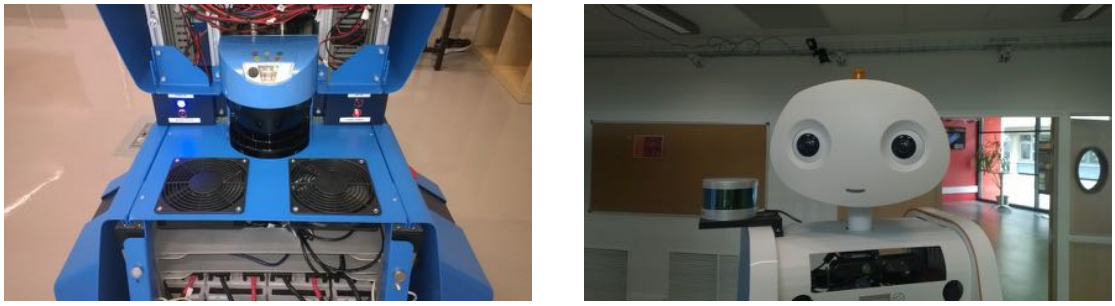


Figure 7: Two fans were added to extract hot air, thereby preventing laptop crashes (left). From the integration week IVb onward support requests were mainly related to the head (right).

2.4.4 Conclusion

Taking apart the additional developments and the improvements, there were 11 support requests regarding the head and two regarding the batteries.

The batteries suffered from a deep discharge, because the platform was not switched off. The batteries could be cured following the procedure of the supplier. It was then decided to add a voltage meter to indicate the battery level and warn users with a strong sound signal before the deep discharge.

The head was clearly the weak point of the platform. It has been designed as specified (2 DoFs). The problem is that the specification was not defined enough, the consortium did not communicate enough to meet the expectation of the ones with the development of the others. Accelerations needed to satisfy interaction with people were much too high for the developed mechanics. In case of follow up of the project, the question of the head should be debated again and a much more robust head should be specified and developed if similar behavior must be realized. The alternative is to decrease again the accelerations of the DoFs.

On the other side, there were no support request regarding the mobile base nor the passenger interfaces (touch screen and boarding card reader).

In conclusion, the technical measures show that the platform is technically robust, with a note for the head that should be improved for the next generation.

3 Component-wise System Evaluations

All major components of the SPENCER platform were evaluated using standard measures that are generally used for the particular components. Comparisons to results on benchmark data sets reported by other research groups were performed. In detail, the following components have been evaluated:

- Mapping and localization (ORU)
- Group- and person detection and tracking (ALU-FR, RWTH)
- Online learning (TUM)
- Close-range perception (RWTH)
- Motion planning (CNRS, ALU-FR)

For a more detailed overview of the components we refer to D6.4.

3.1 Mapping and localization

The mapping component Normal Distributions Transform - Occupancy Map (NDT-OM) has been evaluated in detail in [18]. For the live demo and testing sessions at Schiphol airport, it was difficult to quantify the quality of the map, because we did not have access to ground truth. The quality of the obtained map can be visually assessed by overlaying it with an aerial photo of the airport (see Figure 8). It should be noted that since this environment does not contain loops that could aid global error distribution, but rather consists of long corridors, mapping is very much dependent on precise scan registration in order not to accumulate pose errors that would otherwise show up as bent or shortened corridors. As can be seen in Figure 8, the map is highly metrically correct with respect to the environment, with only a slight bend of the long corridor-like environment of Pier B (far left in the image). This slight bending did not interfere with the quality of localization.

The localization component, Normal Distribution Transform - Monte Carlo Localization (NDT-MCL), has been evaluated in an environment with ground-truth reference localization available [19].

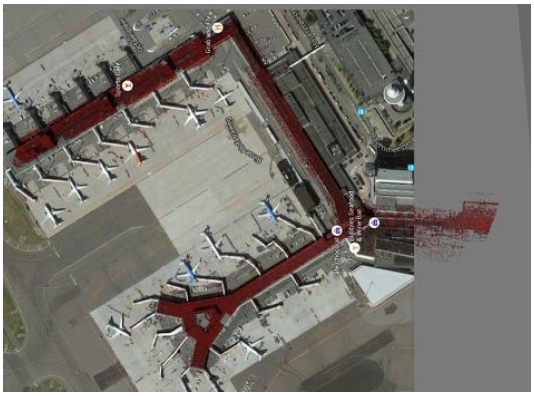


Figure 8: Map built with NDT-OM overlaid with an aerial image of Schiphol airport.

The reported localization error did not exceed 0.07 m. As with mapping, the lack of ground truth at the airport made it impossible to perform quantitative test of localization quality. Instead of this, our measure of success in the live demo at Schiphol airport has been to count how often it was necessary to abort mission execution because of localization issues.

After resolving a technical issue of misalignment of the maps used for planning and localization on March 18 we did not observe any case where it was necessary to abort mission. In other words, from 19 March onward, the robot performed with accurate localization during 18919 m of autonomous operation.

3.2 Group- and person detection and tracking

Person detection and tracking. The person detection and tracking framework of SPENCER has been extensively evaluated in a paper published at ICRA'16 [12]. For this comparison a novel multi-modal evaluation framework has been designed and extensive experiments have been performed on two new challenging datasets. One dataset contains people walking and jumping around the robot in a lab setting, and includes a groundtruth from a motion capture system; the second dataset was recorded in the E pier at Schiphol airport in June 2014 during the first data recording event, and manually annotated.

We compared our multi-modal people tracking system, based upon an extended nearest-neighbor tracker [12, 13], against a tracker from the FP7 EU-project STRANDS [7], an older multi-hypothesis tracker from previous work at ALU-FR [1], as well as the vision-based MDL-tracker from RWTH [9] which was deployed in parallel on the robot to provide bounding boxes for some of the close-range perception modules described in Section 3.4. For these experiments, all tracking systems were provided with the same set of input detections, to prevent varying performance of different detectors from having an effect on the evaluation. The detections used are gathered from different modalities, reaching from upperbody [9] and fullbody detections [20] from the Kinect RGB(-D) data and laser-based leg detections [2]. Some tracking results on the airport dataset are shown in Table 2. It can be seen that the tracking system developed in SPENCER outperforms the other two systems, while being computationally very efficient. As shown in Table 3, it even (slightly) outperforms the vision-based MDL tracking system.

The best results have been achieved by fusing the vision-based upperbody detections in close-range and the laser-based detections especially detecting persons far away. For a more detailed discussion, see [12]. One important remaining issue, which was outside of the scope of SPENCER, was appearance-based person re-identification after lengthy occlusion events, which could significantly reduce the number of relative identity switches (rIDS).

In Linder et al. [12], we also showed that better results (lower false positive rate, thus higher multi-object tracking accuracy (MOTA)) can be achieved if a static occupancy grid map of the environment is available to prevent false tracks from being initiated due to systematic misdetections. While such a static map was not available for evaluation on the airport dataset used in [12], it was available (via NDT-OM) for the final deployment at Schiphol in March 2016, and helped to further improve tracking performance in the real use case.

	Airport Sequence (Pier E) – Multimodal						
Method	MOTA	rIDS	FP%	Miss%	MT	ML	Hz
STRANDS [7]	62.1%	**226**	18.7%	**19.0**%	**114**	27	**6100**
SPENCER [13]	**64.2**%	262	**3.3**%	32.4%	77	33	2222
MHT [1]	60.2%	676	17.2%	22.0%	97	**24**	29

Table 2: Person tracking performance using multi-modal detections (360 degree field of view), up to a distance of 12.0m. The extended NNT developed in SPENCER, shown in the second results row, delivers the best tracking performance in terms of MOTA at low runtime cost.

	Airport Sequence (Pier E) – Front RGB-D						
Method	MOTA	rIDS	FP%	Miss%	MT	ML	Hz
STRANDS [7]	27.7%	227	39.4%	**32.5**%	**92**	**47**	**13701**
SPENCER [13]	**44.4**%	**210**	13.1%	42.1%	63	60	4287
MHT [1]	26.9%	338	39.4%	33.0%	87	51	28
MDL-Tracker [9]	43.7%	428	**12.5**%	43.1%	36	59	53

Table 3: Person tracking performance using only front RGB-D detections (54 deg FOV), to allow comparison with the vision-based MDL tracker which ranks second despite using appearance info.

Group detection and tracking. Group detection and tracking performance is hard to evaluate since the definition of a group can be highly subjective, and task-dependent. In the case of SPENCER, the goal of the robot was to guide groups of passengers to their gate, often encompassing drives through the airport terminal of 400-500m. Some statistics about the estimated number of persons in the guided group of passengers were collected during the end-user evaluation (see Section 4) and are shown in Table 4. However, these tend to be over-estimates caused by project personnel also following the robot (e.g. the two interviewers, the remote emergency stop operator, and an engineer with a laptop), and are therefore not represent a very helpful metric. In practice, no guidance run was aborted due to the robot completely losing track of its group, which might also be due to a 'fail-safe mechanism' implemented in the group guidance supervisor that resorts to just waiting for *any* group to follow the robot, if none of the original person tracks exists anymore.

3.3 Online learning

Our active online object learning method was evaluated on the KITTI benchmark data set consisting of 3D point clouds with pre-segmented object candidates such as pedestrians, cars, and cyclists. To show that our approach can deal with streams of data, we compared it to the situation where the occurrence of objects is uniformly sampled over the time of observation of the objects. Figure 9 shows the resulting learning curves: while standard online learning methods such as the online Random Forest can not deal well with data streams (see left plot), our method uses Mondrian Forests, which can handle this much better (center plot). When used in an active learning scenario, this reduces the amount of required training samples significantly (see left plot). More information our active online object learning method can be found in [15].

#	Day	Time	Actual group size	Initial group size estimate	Final group size estimate	Track	Dist. (m)
1	Sun	12:00	1	9	4	Transfer 2 - Gate B18	461.6
2	Sun	12:40	1	9	2	Transfer 2 - Gate B18	393.7
3	Sun	13:50	1	4	6	Transfer 2 - Gate B18	431.2
4	Wed	13:00	3	6	7	Starbucks - Gate B18	377.4
5	Wed	13:20	3	6	10	Gate B18 - Baggage hall	473.6
6	Wed	13:55	2	7	2	Baggage hall - Gate B18	439.5
7	Wed	15:20	2	5	1	Starbucks - Gate B18	407.1
8	Wed	15:55	1	8	1	Gate B18 - Baggage hall	467.9
9	Wed	16:45	2	5	1	Starbucks - Gate B18[4]	413.5
10	Wed	17:05	2	3	2	Gate B18 - Starbucks	

Table 4: Comparison of actual and estimated group sizes (number of tracked passengers) during end-user evaluations, while the robot was guiding passengers (see Sec. 10). The estimated person count in the tracked group is often higher than the actual value, due to project personnel being close by the robot (e.g. two interviewers, remote e-stop operator, engineer with a laptop)

.

3.4 Close-range perception

Human attribute classification. This module was, in the end, not fully integrated with the final system deployed at Schiphol, as the classifier was trained on higher-resolution data from a Kinect v2 sensor, whereas the robot was (for electrical and mechanical reasons) still equipped with first-generation Asus Xtion sensors. However, an extensive evaluation of frame-by-frame classification accuracy of our novel method for full-body gender recognition in 3D point clouds [14, 11] was carried out for the attributes *gender*, *has long trousers*, *has long sleeves*, *has long hair*, *has jacket* on a previously recorded dataset from a lab environment. Depending on the attribute, we achieve between 75–90% classification accuracy at up to 300 Hz [11], and outperform our own previous baseline and a HOG classifier[5]. Exemplary results are shown in Table 5. For future work after SPENCER, it is planned to evaluate the method on real datasets recorded in the wild and smooth results of the frame-by-frame classifier over time to improve classification accuracy.

Head/body pose estimation. The head and body orientation components (called BiternionNets) were evaluated on publicly available benchmark datasets, where they consistently improved upon the state-of-the-art by relative 7.3% and 6.7% classification accuracy, respectively (see Table 6(top)). When doing regression, BiternionNets reduce the angular error by up to 43.3 degree absolute (see Table 6(bottom)). We also evaluated the effect of the proposed Biternion loss, compared to various simpler baselines (see Table 7). A more detailed experimental evaluation has been published in [5]. As it is unrealistic to annotate exact regression ground-truth for the data recorded at the Schiphol airport, we only evaluated the component qualitatively. The predictions looked reasonable and consistent over time, across many different people that were never seen before.

[5]Note that there are not many existing baseline methods for human attribute classification that operate on RGB-D data.

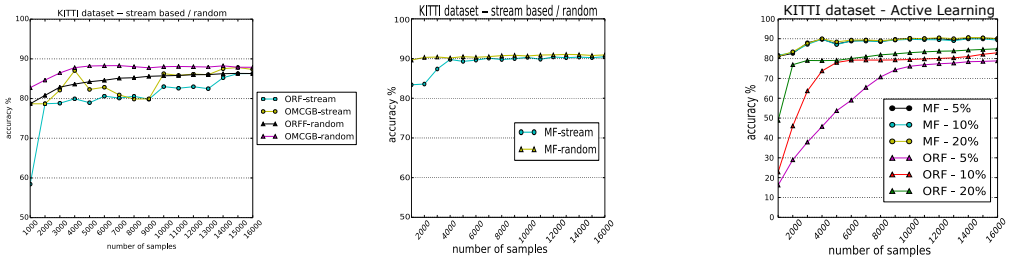


Figure 9: **Left:** Learning curves of two standard online learning methods: online Random Forests (ORF) and online multi-class Gradient Boost (OMCGB), both evaluated on the original and the re-sampled data ("stream" vs. "random"). As we see, the performance of both methods for the original data stream is significantly worse than for the resampled set. **Center:** Learning curves of the Mondrian forest for the "re-sample" experiment. The MF classifier can deal much better with the data stream. **Right:** Classification accuracies for Active Learning using an MF and an ORF, where only 5%, 10% and 20% of the most uncertain data points are queried. Again, the MF clearly outperform the standard ORF.

Gender	(1)	(1)–(3)	(1)–(4)
HOG	78.0%	76.9%	77.0%
[14]	89.8%	83.7%	82.6%
Ours	**90.4%**	**87.0%**	**86.3%**

Long trousers	(1)	(1)–(3)	(1)–(4)
HOG	65.0%	60.0%	59.4%
[14]	69.4%	66.0%	64.1%
Ours	**83.6%**	**78.0%**	**76.2%**

Long sleeves	(1)	(1)–(3)	(1)–(4)
HOG	63.2%	60.8%	60.7%
[14]	62.3%	61.8%	61.0%
Ours	**76.9%**	**73.8%**	**72.8%**

Long hair	(1)	(1)–(3)	(1)–(4)
HOG	74.3%	72.6%	72.7%
[14]	83.7%	77.9%	77.2%
Ours	**87.2%**	**83.3%**	**82.9%**

Table 5: Classification accuracy of different human attributes on ALU-FR's human attribute dataset, including static poses (seq. 1), walking sequences (seq. 2+3) and close-up interaction (seq. 4) [11].

Articulated upper-body pose estimation. For this task, we developed approaches in two directions. The RGB-D based skeleton tracker, used for the Schipol deployment and described in D3.3, was heavily optimized for run-time and low computational effort in order to run on the robot in parallel to all other perception components. Qualitative tests showed that the tracker ran at video frame rate and worked well for mostly front-facing walking people in close range to SPENCER robot. As the tracker was trained with scenario-specific training data, however, its performance could not be formally evaluated on benchmark datasets from the literature.

In parallel, we developed a Convolutional Neural Network (CNN) based body pose estimation approach that operates on single color images. This approach could be evaluated on the standard human body pose estimation benchmarks, which we did to great success. As shown in a recent BMVC'16 submission [17], our approach achieves top results on the MPII benchmark (see Tab. 8) and close-to-top performance on the LSP and FLIC datasets.

	HIIT	HOCoffee	HOC	QMUL	
# Samples	12 000/12 007	9522/8595	6860/5021	7603/7618	9813/8725
# Classes	6	6	4	4	4 + 1
Tosato *et al.* [22]	96.5%	81.0%	78.69%	94.25%	91.18%
Lallemand *et al.* [10]	-	-	79.9%	-	-
Our CNN	**98.70%**	**86.99%**	**83.97%**	**95.58%**	**94.30%**

	IDIAP Head Pose			CAVIAR-c	CAVIAR-o
# Samples	42 304/23 991			10 660/10 665	10 802/10 889
Pose range	pan	tilt	roll	pan	pan
	[-101,101]	[-73,23]	[-46,65]	[0, 360]	[0, 360]
Tosato *et al.* [22]	$10.3°\pm10.6°$	$4.5°\pm5.3°$	$4.3°\pm3.8°$	$22.7°\pm18.4°$	$35.3°\pm24.6°$
Ba & Odobez [3]	$8.7°\pm9.1°$	$19.1°\pm15.4°$	$9.7°\pm7.1°$	-	-
Our CNN	$\mathbf{5.9°}\pm\mathbf{7.2°}$	$\mathbf{2.8°}\pm\mathbf{2.6°}$	$\mathbf{3.5°}\pm\mathbf{3.9°}$	$\mathbf{19.2°}\pm\mathbf{24.2°}$	$\mathbf{25.2°}\pm\mathbf{26.4°}$

Table 6: **Head/body pose estimation results:** (top) Class-average accuracies on the four head pose classification datasets from [22]. The sample counts refer to the provided train/test splits. (bottom) A comparison to two pose regression datasets from [22]. The first number is the mean absolute angular deviation, the second its standard deviation across test-samples. Our CNN approach obtains state-of-the-art results on all datasets. (Tables from [5])

Method	MAE
Linear Regression	$64.1°\pm45.0°$
Naive Regression	$38.9°\pm40.7°$
Von Mises	$29.4°\pm31.3°$
Biternion	$21.6°\pm25.2°$
Biternion+Von Mises	$\mathbf{20.8°}\pm\mathbf{24.7°}$
Benfold&Reid [4]	$25.6°$ / $64.9°$

Table 7: **Head/body pose estimation:** Improvement in Mean Angular Error (MTE) our Biternion CNN achieves, compared to various baselines on the TownCentre dataset [4] (Table from [5]).

3.5 Motion planning

During integration week V we have further fine-tuned the motion planner to take the particular dynamics of Schiphol Airport into account. As a measure of success of the motion planner we compared the distance the robot drove autonomous compared with not autonomous (in meters).

Table 9 details the distances traveled every day, which is reported as sum of the distances traveled by the robot autonomously and not autonomously. Overall the robot drove autonomously circa 46 kilometers in 15 days. Only on particular days like the one of the data recording and the mapping sessions, the robot did not drive autonomously.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	PCKh
Hu et al [8]	95.0	91.6	83	76.6	81.9	74.5	69.5	82.4
Carreira et al [6]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al [21]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Pischulin et al [16]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Wei et al [23] + *LSP*	*97.8*	*95.0*	*88.7*	*84.0*	*88.4*	*82.8*	*79.4*	*88.5*
Ours	**96.8**	**93.1**	**85.2**	**79.7**	**85.8**	**78.7**	**71.8**	**85.1**

Table 8: Comparison of our CNN-based body pose estimation approach with recent state-of-the-art methods on the MPII benchmark using PCKh @ **0.5**. Note that the approach by Wei et al. [23] used an extended training set and is therefore not directly comparable (Table from [17]).

4 End-user evaluations

The research and integration activities over the past three years have resulted in a platform which is technically safe and capable of navigation autonomously at an airport. One of the moments of trusts of this project is the user evaluations conducted with the SPENCER platform, to prove that besides being technically capable, a robot like SPENCER is also deemed acceptable and useful by representative users, in this case airport passengers.

Two user evaluation studies have been conducted with the SPENCER robot to assess user experience of the SPENCER robot. The first event took place in a semi-public space at the University of Freiburg, where we assessed participants impression of the robot. Questionnaire data showed that participants were in general positive about the robot, and the open questions revealed that participants were in general positive about the autonomous navigation and obstacle avoidance capabilities of the robot. Furthermore, most participants believed the robot acted in a socially normative way.

The second user study has been conducted at Schiphol Airport during the integration & demonstration event in March 2016. Over the course of two days ten groups of users have been invited to be guided by the SPENCER robot, and afterwards participate in a brief interview. In general, the user groups can be characterized as diverse. In this section we will describe the method and results of the user study conducted at Schiphol Airport. All participants participated in one of two related test scenarios, which we will describe in Section 4.1. After this we will describe the sample (Section 4.2), measures (Section 4.3) and results (Section 4.5).

4.1 Test scenarios at Schiphol

The original test scenario as defined during the use-case workshop held on April 9, 2013 included SPENCER awaiting passengers at their arrival gate (in the non-Schengen area). Due to airport refurbishments the consortium modified the test scenario into the current scenario, which allows for the testing of all components, the main difference is that the distance over which is being guided is smaller, and the robot does not collect passengers itself.

Tests at Schiphol Airport took place in the Schengen-part of the terminal: this area encompasses Lounge-1, Transfer-2, the B- and C-piers and the upper level of the D-piers. As SPENCER would

Date	Dist. Auton.	Dist. Not Auton.	Dist. per day	Day's event
09/03/16	0	4210	4210	Mapping session
10/03/16	589	1055	1644	
11/03/16	2168	2701	4868	
12/03/16	371	1555	1926	Mapping session
13/03/16	3475	1099	4574	
14/03/16	3932	2176	6107	
15/03/16	3768	882	4650	
16/03/16	2248	1739	3986	
17/03/16	3830	2249	6079	
18/03/16	7060	1877	8937	Film crew
19/03/16	2235	1096	3331	
20/03/16	4397	627	5024	User study (morning)
21/03/16	4813	1084	5897	
22/03/16	3562	1379	4941	Review
23/03/16	3912	1980	5891	Film crew and user study
24/03/16	0	1051	1051	Data Recording and Packing
Total	46360	26759	73118	

Table 9: The table details the distances (in meters) traveled every day (**Dist. per day**) during the final demo in Schiphol. The total distance is split into two: the distance traveled with the robot driving autonomously (**Dist. Auton.**) and not autonomously (**Dist. Not Auton.**). Only on particular days like the one of the data recording and the mapping sessions, the robot did not drive autonomously, in the remaining days the robot was mainly operated autonomously.

have to take an elevator to reach the Schengen D-piers this area was deemed beyond the scope of the current tests, and therefore not mapped.

In the first test scenario SPENCER would pick up passengers in Lounge-1 or at the Starbucks, and guide them to gate B18 (Figure 10). Gate B18 was chosen as it would require SPENCER to pass passengers waiting at other gates, and a number of shops along the way. This scenario required passengers to interact with the robot using a test boarding card. The second scenario which we evaluated was one where SPENCER picked up passengers in the B-piers and guided them to Lounge-1, specifically the stairs to the baggage hall (Figure 11). In this scenario passengers did not have to use a test boarding card, rather they could select this point on the touchscreen display.

In each scenario SPENCER guided through the more crowded Lounge-1, the junction near Transfer-2 and the corridor between B-piers and Lounge-1, which can be characterized as a long hallway, featuring moving walkways and static objects such as columns, benches and artworks (Figure 12).

4.2 Participants

Ten trials with 18 participants were conducted over two days (Sunday 20 March and Wednesday 23 March). The difference between these trials lies in the measures; the questionnaire on Wednesday was slightly modified; we will elaborate on this difference in Section 4.3). Information about the

#	Day	Time	Group size	Track	Distance (m)	Duration (mm:ss)	Comments
1	Sunday	12:00	1	Transfer 2 - Gate B18	461.6	07:36	
2	Sunday	12:40	1	Transfer 2 - Gate B18	393.7	08:05	
3	Sunday	13:50	1	Transfer 2 - Gate B18	431.2	14:41	
4	Wednesday	13:00	3	Starbucks - Gate B18	377.4	06:35	Project staff
5	Wednesday	13:20	3	Gate B18 - Baggage hall	473.6	10:17	
6	Wednesday	13:55	2	Baggage hall - Gate B18	439.5	16:19	
7	Wednesday	15:20	2	Starbucks - Gate B18	407.1	06:44	
8	Wednesday	15:55	1	Gate B18 - Baggage hall	467.9	08:09	
9	Wednesday	16:45	2	Starbucks - Gate B18[6]	413.5	06:33	
10	Wednesday	17:05	2	Gate B18 - Starbucks	-	-	KLM staff

Table 10: 18 participants divided over 10 groups, participated in the user evaluation at Schiphol. Data from trial #10 missing due to a recording error.

group distribution is provided in Table 10.

The sample consisted of 11 males and 7 females, aged between 26 and 54 (M=37.06, SD=9.04). Ten participants indicated the purpose of their journey was business, 8 participants indicated pleasure.

4.3 Measures

Three types of measures were collected during the user studies, these being:

1. Feedback questionnaire (individual, self-reported)
2. Interview (group)
3. Notes taken by one of the researchers during guiding[7]

[7] We added this measure for the user studies on Wednesday, as we found out on Sunday that because the distance traveled was quite long, participants forgot specific positive or negative events which occurred during the guiding.

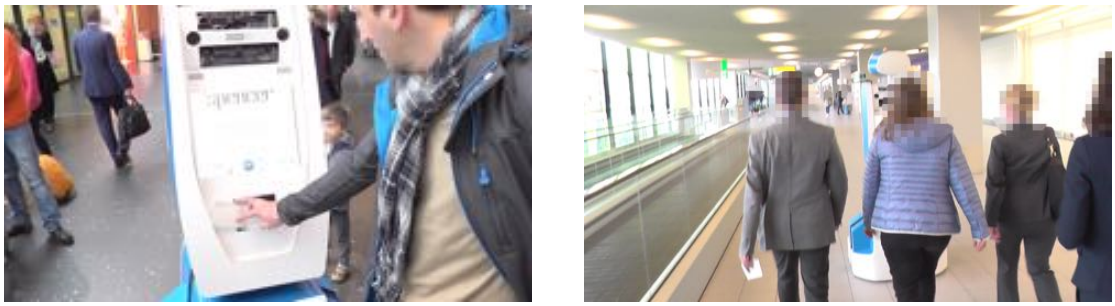


Figure 10: The first scenario required passengers to identify themselves using their test boarding card (left), after which they were being guided to gate B18 (right)

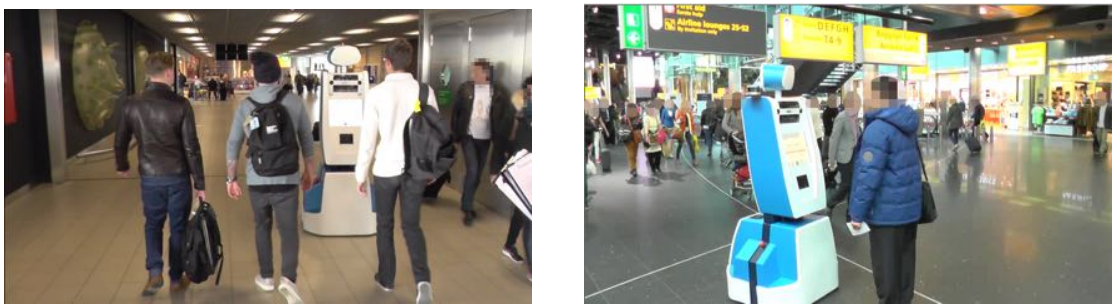


Figure 11: In the second scenario the robot guided passengers from the B-piers towards the baggage hall in Lounge-1

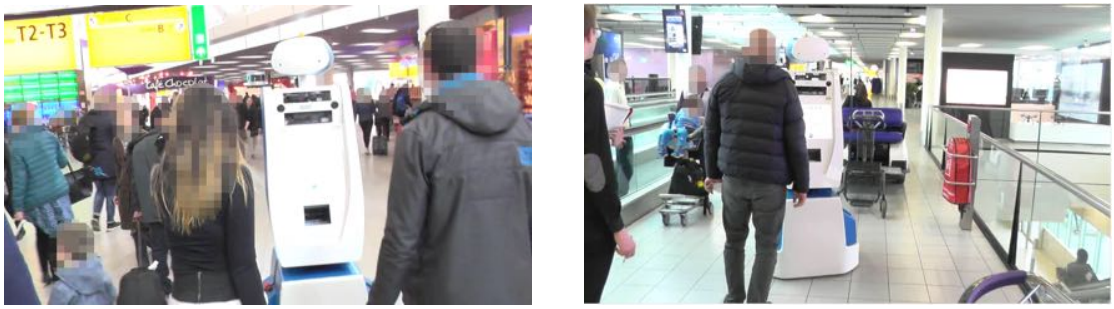


Figure 12: SPENCER navigating through the junction behind Transfer-2 (left) and waiting for another passenger to pass (right)

We will discuss each of the three measures briefly below.

The feedback questionnaire consisted of 3 journey related questions (start, end and purpose), the group size and participants' gender and age. The remainder of the questionnaire consisted of 8 7-point Likert scaled statements about the appearance- and behavior of the robot, and the influence of the service on participants' customer satisfaction. The questionnaire on Wednesday included an additional 2 items about the robot, as well as a question on participants' opinion of robots in general. Both questionnaires can be found in Appendix B.

Semi-structured interview were conducted with the groups, addressing topics such as the first impression of the robot, positive & negative experiences, improvement of customer satisfaction and general points of improvement. Interviews lasted between 3 and 19 minutes.

On Wednesday one of the experimenters accompanied the group, and made notes about events participants found noteworthy. Before the experiment, participants were encouraged to think aloud about their experiences.

4.4 Data analysis

For the purpose of this analysis, we report the results of the feedback questionnaire and interviews. The feedback questionnaire was analyzed using standard statistics software. The interviews with the recruited participants have been transcribed on a detailed level. In the next step, these data were coded

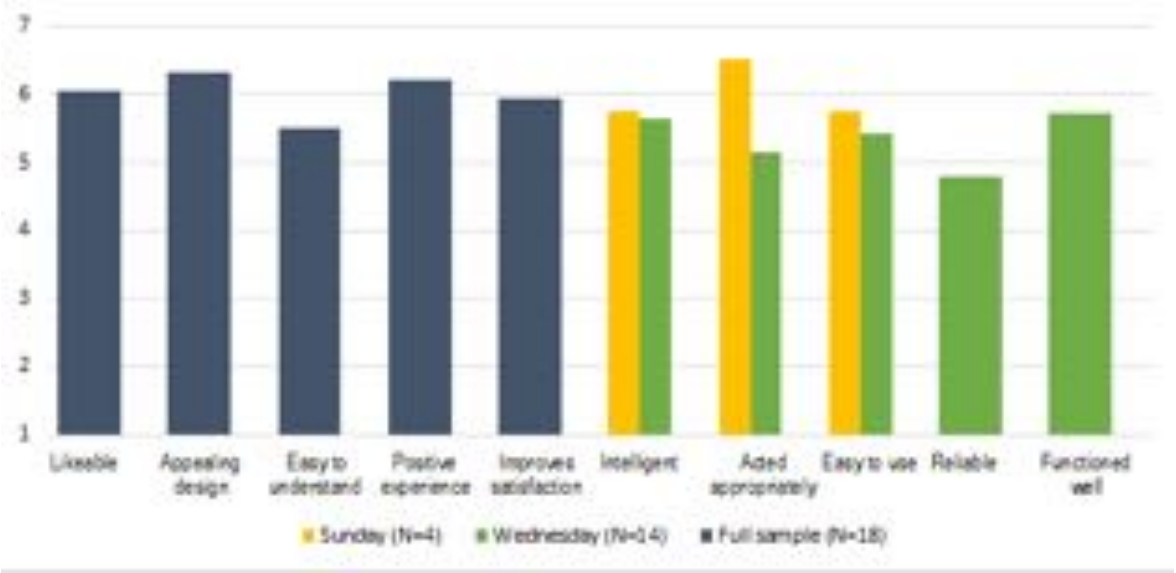


Figure 13: Results of the self-reported questionnaire indicated participants were in general satisfied with SPENCER's performance.

using a qualitative data analysis software package[8], which with each statement made by a participant was coded in two iterations. In the final step, all statements having a particular code were analyzed, and the research team searched for commonalities and noteworthy experiences.

4.5 Results

In this section we will present the results of the end-user evaluation. We will first present general findings, as gathered through the feedback questionnaire. In the consecutive subsections we will discuss particular topics discussed during the interviews. Where applicable we will illustrate our findings with quotes and photos.

Figure 13 shows the results of the feedback questionnaire. Two questions were reformulated on the "Wednesday questionnaire"', therefore we present means for both questions. As can be seen in Figure 13 self-reported scores are generally high for all measures; this warrants the first conclusion participants in general had a positive impression of the robot. The items "reliability" and "acted appropriately" scored relatively low; the interview data allows us to interpret this result, which we discuss further on in this section.

Though the sample size is small - the findings are therefore not generalizable - we have conducted T-tests to test for differences on the dependent variables. Specifically we have tested for differences in *gender* and *journey purpose*. We found a significant difference of gender on likeability and appearance: females thought the robot was more likeable (M=7.00, SD=0) than men (M=5.45, SD=1.51), $t(16)$=-2.682, $p<0.05$. In a similar fashion females thought the robot's appearance was significantly more appealing (M=7.00, SD=0) compared with males (M=5.91, SD=0.831), $t(16)$=-3.434, $p<0.01$.

[8] Atlas.ti

The interview data has been analyzed using the approaching outlined in Section 4.4.

4.5.1 General impression

In general participants had a positive impression of the robot. Ten out of fifteen participants who commented on this question indicated their first impression was that the robot was "nice","great", or "helpful"; positive in general. Other participants were surprised to see a robot at the airport, or could not fully identify the function of the robot. One participant commented that

"My first impression is that I thought it was a security robot, so I didn't think it was a guided, but because of the color looks like security, police, something."

We will further comments on the appearance of the robot in Section 4.5.2. Another feature which surfaced regularly was the user interface.

To guide our semi-structured interview, we designed an interview covering various themes we believed were important, such as the appearance and behavior of the robot, ease of use, whether they received enough feedback from the robot, and if there were situations in which participants believed the robot was especially rude or polite. We also discussed in which scenario's participants thought this robot would be especially useful. Before going into those topics, we first asked participants if they could name particular positive and negative aspects of their experience with the robot. We have incorporated these comments into the various sections.

4.5.2 Appearance

Eight participants commented specifically on the appearance of the robot. Six participants positive about the robot, as illustrated by the quote above. One participants voiced his concern about the two emergency stop buttons, as he associated that with danger, and another participant said the colors first made him think that it was a security robot, because of the blue color. The colors were indeed not always immediately associated with the KLM brand colors, as evidenced by the following quote:

"Maybe do a girl one, make it pink. For like, political correctness. Obviously you'll get someone who says like 'why Spencer'. I don't know. I'm just throwing it."

4.5.3 Guiding behavior and speed

Fourteen participants commented on the speed of the robot; three participants who followed the robot indicated the robot could have driven faster, however, these participants followed the robot when it was not very crowded in the terminal. Five participants indicated the robot drove too fast, especially if the robot were to guide a family around.

Two recurring comments regarded the general driving behavior of the robot: participants liked the fact that the robot stopped when people were too close to the robot, however, especially in more crowded situations this happened too often. Coupled with the fact that replanning was perceived as taking quite long, participants general impression was that the robot was less suited for guiding passengers in a congested area, especially if they were under time pressure.

"Yeah, a slower process. Like instead of just stopping, just gradually come to a halt, I think that would be better, yeah."

4.5.4 Appropriate behavior

We asked people if there were particular moments where they felt the robot behaved in particular rude or polite. Answers were quite diverse, and ranged from neutral to "all good". The tone of the answers was similar as we found with the feedback questionnaire: moderately positive. One participant who was guided commented that: *"Well, here it is really really crowded. So this is something like, there were a few moments when you were just running, and some moments it was so slow. You don't actually have an average, because it was too much of both sides. And also the stops were really like... woof."*

Another participant commented that it would perhaps be beneficial to have specific lines for the robot, as the robot stopped quite often in crowded environments. This is one of the recurring themes which we got from the interviews, and we will reflect on this when discussing suggested improvements (Section).

4.5.5 Reliable & Easy to use

Seven participants were explicitly asked whether they thought the robot was reliable. All participants indicated the robot was reliable. Five passengers commented on the easy-of-use, and indicated it was very easy, though at the same time all participants who were guided towards the B-piers, and had to use a test boarding card, had issues using the boarding card reader. All passengers commented on this issue.

"Yes, except for, I don't know how, in which direction to scan those codes. I only worked when I hold it like that."

4.5.6 Feedback

With the exception of the use of the boarding card reader, participants believed there was enough feedback while being guided. All participants who had to use their boarding card reader commented that it was difficult to understand the workings of the boarding card reader. Furthermore, participants suggested to add a map to the user interface, and perhaps additional information, in particular where to collect their luggage (in case of being guided to the baggage hall), and the location of facilities such as toilet and shops.

"So while we were being guided, the feedback was definitely enough, so the information it was showing how closing we were to the gate, and everything, and it was great. I think that at the beginning the phase where I was supposed to scan the boarding pass, there could be more explanation on how to do it, just to put the boarding pass and wait"

Eight participants commented that feedback could be improved by providing more auditory feedback. Current auditory feedback was limited to "Excuse me", when the robot was blocked by people. Both auditory feedback to warn bystanders, and to inform the guided passengers about the process (e.g. "Let's go", "we're almost there"). None of the interview participant expected the robot to be

able to conduct complex conversations, as illustrated by this comment: *"There could be, further in the project, there could be more interaction with the robot. So not only through the interface, but maybe also some voice-over, something like that."*

4.5.7 Improving customer satisfaction

We asked 13 people if a service like SPENCER would improve participant's customer satisfaction; 12 people answered positively to this question, one participant was undecided. Three participants stated that it was a good thing that KLM invested in new technologies. Other participants also indicated that this service at Schiphol would be more for entertainment, as it wasn't too hard to find your way at Schiphol - for experienced travelers.

"I think it shows that KLM is not afraid of technology itself."

4.5.8 When using SPENCER would be attractive

In the interviews on Wednesday we asked participants in which scenarios a service like SPENCER would be useful. Three out of six participants indicated they would not need the robot, either because they were familiar with the airport, but mainly because they believed the robot would be more useful for people not accustomed to flying (elderly) or with disabilities. When asked if they would find SPENCER useful at an unknown airport they did indicate SPENCER would be useful.

Two participants believed SPENCER would be especially useful in the outer limits of the airport. One participant had a recent experience at an airport abroad where she had missed her flight, and did not receive any information. She indicated this would be an excellent use case for SPENCER as it could tell passengers where to go, and even guide them to (f.e.) a transfer desk.

"No, we're talking about Schiphol, and you have to understand that it's one of best organized airports in the world. Which means that for the people it is really easy to understand where to go, it's not like Heathrow or any other airport [...] So here it's really easy, so perhaps the robot should give more service in order to be sustainable."

4.5.9 Improvements

Additional improvements highlighted by the participants include extra guiding services, such as the ability of SPENCER to guide you to toilets and specific restaurants. Two participants mentioned that a robot capable of carrying hand luggage would be convenient as well.

Five participants commented on the challenges of operating a robot in an environment as crowded as Schiphol: three participants explicitly mentioned the stopping behavior should be improved, as the robot currently stopped too sudden when encountering obstacles. Two participants suggested that it might be better to test in a less-congested area.

Finally, two participants recommended that the function of the robot was somewhat unclear, and that a dedicated place where the robot would be when idle (available for services) could be an improvement if and when SPENCER would be deployed.

4.5.10 Summary of recommendations, lessons learned

In this section we have described results of a user study using both qualitative and quantitative data. We have clustered answers from the interviews and drawn general conclusions related to several themes, ranging from the behavior of the robot to usage scenario's for robotic services at airports. To summarize the findings from the user evaluations:

- Participants were generally happy with SPENCER's guiding performance
- SPENCER can be described as friendly-looking, easy-to-use and reliable
- Shortcomings of the tested demonstrator robot are primarily related to the user interface (boarding card reader), and the abrupt stopping of the robot - the later as a safety precaution
- Participants frequently mentioned SPENCER being useful for people inexperienced with flying, or new to the particular airport
- Participants were of the impression that SPENCER improved their customer satisfaction
- Main improvements for SPENCER include technical improvements (stopping less for obstacles) and extended use case improvements (such as guiding to shops and toilets)

5 Conclusion

In this report we have described the evaluation of SPENCER's deployment at Schiphol airport with respect to technical and user experience measures. During these two weeks at Schiphol, Spencer drove 46 kilometers autonomously without major technical problems.

The technical components all worked well enough to conduct tests with real passengers in a guiding scenario. With the exception of the abrupt stopping and the time it took to re-plan a path, no technical failures occurred during the user studies. In general the participants were satisfied with the performance of SPENCER.

During the project lifetime there were delays, for example to guarantee platform safety. These delays led to a less elaborate use case, which was an aspect of the experience participants indicated could be improved.

The user tests revealed that guiding by itself does not necessarily add to the customer experience of traveler familiar with Schiphol Airport, rather that these travelers would appreciate additional services which would technically be feasible but more content-specific. Examples include participants who requested more information on (f.e.) shops, or a better user interface for the boarding card reader are all aspects which were not specifically described in the original project plan, but turned out to be important for the total user experience.

In general we conclude that we have succeeded in deploying a robotic demonstrator at Schiphol Airport for the purposes of guiding transfer passengers. The robot worked within parameters as could be expected for a demonstrator robot, and the user studies as conducted gave us new insights for research on domestic service robots.

References

[1] K. O. Arras, S. Grzonka, M. Luber, and W. Burgard. Efficient people tracking in laser range data using a multi-hypothesis leg-tracker with adaptive occlusion probabilities. 2008.

[2] K. O. Arras, O. Martínez Mozos, and W. Burgard. Using boosted features for the detection of people in 2d range data. 2007.

[3] S. O. Ba and J.-M. Odobez. Evaluation of Multiple Cue Head Pose Estimation Algorithms in Natural Environements. 2005.

[4] B. Benfold and I. Reid. Unsupervised Learning of a Scene-Specific Coarse Gaze Estimator. In *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[5] L. Beyer, A. Hermans, and B. Leibe. Biternion nets: Continuous head pose regression from discrete training labels. In *Pattern Recognition, Proceedings of GCPR 2015*, volume 9358 of *Lecture Notes in Computer Science*, pages 157–168. Springer, 2015.

[6] J. Carreira, P. Agarwal, K. Fragkiadaki, and J. Malik. Human Pose Estimation with Iterative Error Feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] C. Dondrup, N. Bellotto, F. Jovan, and M. Hanheide. Real-time multisensor people tracking for human-robot spatial interaction. In *Workshop on Machine Learning for Social Robotics at International Conference on Robotics and Automation (ICRA)*, 2015.

[8] P. Hu and D. Ramanan. Bottom Up and Top Down Reasoning with Hierarchical Rectified Gaussians. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] O. H. Jafari, D. Mitzel, and B. Leibe. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. 2014.

[10] J. Lallemand, A. Ronge, M. Szczot, and S. Ilic. Pedestrian Orientation Estimation. In *German Conference on Pattern Recognition (GCPR)*, 2014.

[11] T. Linder and K. O. Arras. Real-time full-body human attribute classification in rgb-d using a tessellation boosting approach. 2015.

[12] T. Linder, S. Breuers, B. Leibe, and K. O. Arras. On multi-modal people tracking from mobile platforms in very crowded and dynamic environments. In *ICRA*. 2016.

[13] T. Linder, F. Girrbach, and K. O. Arras. Towards a robust people tracking framework for service robots in crowded, dynamic environments. In *Assistance and Service Robotics Workshop (ASROB-15) at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015.

[14] T. Linder, S. Wehner, and K. O. Arras. Real-time full-body human gender recognition in (RGB)-D data. 2015.

[15] A. Narr, R. Triebel, and D. Cremers. Stream-based active learning for efficient and adaptive classification of 3d objects. In *Int. Conf. on Robotics and Automation*, May 2016.

[16] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. DeepCut : Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] U. Rafi, B. Leibe, and J. Gall. An Efficient Convolutional Network for Human Pose Estimation. In *submitted to BMVC*, 2016.

[18] J. Saarinen, H. Andreasson, T. Stoyanov, J. Ala-Luhtala, and A. J. Lilienthal. Normal distributions transform occupancy maps: Application to large-scale online 3D mapping. pages 2233–2238. IEEE, 2013.

[19] J. Saarinen, H. Andreasson, T. Stoyanov, and A. J. Lilienthal. Normal distributions transform Monte-Carlo localization (NDT-MCL). In *IEEE Int. Conf. Intell. Robot. Syst.*, pages 382–389, 2013.

[20] P. Sudowe and B. Leibe. Efficient use of geometric constraints for sliding-window object detection in video. In *Computer Vision Systems*. 2011.

[21] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient Object Localization Using Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] D. Tosato, M. Spera, M. Cristani, and V. Murino. Characterizing Humans on Riemannian Manifolds. 35(8):1972–1984, 2013.

[23] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Appendix A: User study with the Spencer platform in Freiburg

The influence of a head turn on subjective- and objective evaluation of a guide robot

Michiel Joosse
Human Media Interaction Group
University of Twente
Enschede, the Netherlands
m.p.joosse@utwente.nl

Luigi Palmieri, Billy Okal, Timm Linder, Kai Oliver Arras
Social Robotics Lab, dep. of Computer Science
University of Freiburg
Freiburg, Germany
palmieri, okal, linder, arras@cs.uni-freiburg.de

Vanessa Evers
Human Media Interaction Group
University of Twente
Enschede, the Netherlands
v.evers@utwente.nl

ABSTRACT

In this paper we present a study where small groups of people (N=69) followed the SPENCER robot for a guided tour through a semi-public space. We manipulated the head direction behavior of the robot. Our results show that participants were satisfied with the appearance and behavior of the SPENCER robot. The manipulation however did not yield any significant differences.

CCS Concepts

•**Human-centered computing** → *Empirical studies in HCI;*

Keywords

Guide robot, Spencer, Human-Robot Interaction, Groups

1. INTRODUCTION

Social robots, and robots in general, come in all kinds of forms and shapes. A social robot, operating autonomous or not, differs from a "regular" robot in they operate in an environmental specifically designed for humans. One feature many social robots have in common is that they have a design which evokes certain anthropomorphism, e.g. people attribute humanlike characteristics to the robot [6].

The SPENCER robot is an autonomous guide robot specifically designed to provide services to transfer passengers at international airports [21]. SPENCER has been designed to be anthroporphised, e.g. it has a human-like head, with eyes. As can be seen in Figure 1 the design features humanoid elements, though it is not so humanoid that SPENCER elicits too high expectations from users, nor that the appearance is that humanoid that is would seem uncanny, e.g. the uncanny valley theory [17].

ACM ISBN . . . $15.00

DOI:

In our previous work we have investigated which direction the head of a human robot should face when guiding a small group of people, from this pilot we found a strong preference for a forward-facing head [11]. However, that work was limited in that the anthropomorphic design was very limited, and that hardware restrictions influenced the overall evaluation of the robot.

The contribution of the research presented in this paper is two-fold. First, we evaluate the SPENCER platform with respect to it's appearance and basic behavior capabilities in a guiding task. This is the first evaluation conducted with novice users. The second contribution of this research is that we investigate to what extend the behavior of the head during guiding results in a different evaluation of the robot as found in our previous research.

In order to do so, we will first discuss related work (Section 2), followed by presenting a pilot study in which we tested which head turn behavior is most appropriate. This head turn was then implemented on the SPENCER platform, and evaluated in a between-groups experiment. We will present the method (Section 4) and initial results (Section 5) of this experiment. We will conclude with a discussion of the results found (Section 6).

Figure 1: The SPENCER robot

2. RELATED WORK

In this section we will discuss related work on acceptance of social robots (Section 2.1) and non-verbal communication with robots (Section 2.2). We will end this section with the hypotheses governing the design of the current study.

2.1 Acceptance of social robots

Fong et al. [7] distinguishes social robots from "normal robots" by emphasizing on the fact that social interaction plays a key role; they can autonomously interact with humans in a socially meaningful way. The following definition, adapted from Bartneck & Forlizzi [2], will be used for social robot:

"A (social) robot is a physically embodied machine which is specifically designed to operate in a human environment, and interact with humans, in ways deemed appropriate by those humans the robot is intended to work with."

The acceptance of social robots, and the factors influencing acceptance, have been studied in HRI literature. Examples include the work by Heerink [9], de Graaf [8] and Beer et al. [4]. All these works identify (and test) possible variables which influence the acceptance, or use intention, of social robots. The work of Beer et al. [4] is a more general literature review with the goal of identifying factors which predict acceptance of robots, whereas the works of Heerink [9] and de Graaf [8] are more specifically focused on robots in domestic environments. Beer et al. [4] defined three categories of variables which potentially influence acceptance of robots, these being function (control method, autonomy level), social capability (social intelligence, emotion expression and non-verbal social cues), and appearance (human likeness, form). Specifically relevant for the SPENCER project, de Graaf [8] found social norms to be the core factor in the proposed social robots acceptance model.

According to Breazeal [5] "the robot's observable behavior and the manner in which it responds and reacts to people profoundly shapes the interaction and the mental model people have for the robot".

Based upon our proposed model we believe that the acceptance of social robots (or the evaluation of specific behavior) is dependent on several factors, most importantly the appearance of the robot, which is mostly static, and can be evaluated by examining the anthropomorphic design of the robot in terms such as the perceived human likeness and perceived anthropomorphism. The second major factor of influence is the perceived social competence of the robot; the perception of the actions of the robot.

2.2 Non-verbal communication with gaze

People communicate all the time. Merriam-Webster [16] defines communication as "a process by which information is exchanged between individuals through a common system of symbols, signs, or behavior", and while this also includes language, there are more communication channels people frequently use. Norris [19] poses that "all movements, all noises and all material objects carry interactional meaning as soon as they are perceived by a person", under which she clusters both verbal and non-verbal behaviors, but also images. For Human-Robot Interaction (HRI) this implies that also the appearance of a robot; for example the head, communicates a certain message to people even before they have started interacting by means of vocal or touch input.

In social-psychology, research has been conducted into eye contact, and among other the functions of maintaining eye contact. Argyle & Dean [1] summarize these functions, which includes among others signaling that the channel is open, information-seeking, and establishing and recognizing a social relationship. Furthermore, it has been shown to impact our image of others, and whether positive or negative, this being a sign of potential social interaction [22].

Different types of social robots exist, and some robots have heads and eyes as well. These eyes and heads are not necessarily used for navigation and interaction as robots possess other ways of sensing their environment. Eyes, and gazing with those eyes, have however been shown to impact people's reaction to robots, and guide robots in particular. Karreman et al. [13] observed that people unconsciously perceive cameras as eyes on robots, and when oriented at the point of interest (instead of gazing at the guides) participants stood closer to the robot. Similarly, when providing explanations of artwork gazing at participants increases their evaluation of the robot, though it did not improve recall of information [12].

Robots can use non-anthropomorphic cues in different ways than humans, e.g. in the guiding context they can display route information rather than eyes. On the other hand, literature indicates that people use a combination of head and eye movement to non-verbally indicate their walking direction [10] and users might expect robots to do the same; thus directing their gaze ahead of them.

In HRI, two studies have investigated the direction of the head while guiding, or driving to attract attention of people. Shiomi et al. [20] conducted an experiment with the Robovie robot which drove either forward or backward while guiding participants in a mall (over a short distance). More bystanders joined when the robot moved backwards compared with frontwards, and that more people were inclined to follow the robot the entire time when moving backwards. In our own work [11], we asked people to follow a robot with its head either directed to the front and back, and we found a significant preference for a head turned forward. Unfortunately both studies featured a different context (lab vs. shopping mall), participants of two different cultures (Asian vs. European). Furthermore, while we have an idea which head direction people prefer from a guiding robot, we do not know yet if such a robot would be perceived having a different mental model, e.g. more or less competent.

2.3 Hypotheses

For this study we have the following hypotheses:

H1. The operationalization of the socially normative behavior will be evaluated differently; in particular participants will perceive the socially normative robot as being significantly more socially competent.

H2. Evaluation of the robot's social competence is not dependent on socially normative behavior, rather there is a correlation with attitude towards robots in general.

H3. Evaluation of the robot's social competence is not dependent on socially normative behavior, rather there is a correlation with the evaluation of the anthropomorphic design of the robot.

In order to test these hypotheses we will first conduct a pilot study, which will result in one head turn behavior for the SPENCER robot which is consider to be "normal". This behavior will then be implemented on the robot, and in the subsequent experiment we will test our hypotheses.

3. PILOT: PROTOTYPING A HEAD TURN FOR SPENCER

In this section we present a pilot study. In our previous work we found evidence that people following a robot have a preference for a guide robot which looks ahead of them when guiding [11]. One possible explanation could be that people think the robot should face it's driving direction, an alternative explanation is that it might feel creepy to have a robot driving backwards in general.

To select the most appropriate head turn for the SPENCER robot we designed 16 head turns, which varied among two "axis": the type of the turn, and the angle of the turn, all of which we will describe in Section. Inspired by, we created short video sequences of the head turns which we distributed through the CrowdFlower platform[1].

3.1 Stimuli

We designed five distinct turn behaviors; which differed in the number of times the head would stop during a turn (Figure 2). For example, in sequence A, the head would first turn 180 degrees counterclockwise, stop, turn 15 degrees counterclockwise, stop, turn 30 degrees clockwise, stop, and turn back to the starting point. Other sequences, such as sequence D, would only stop at one point.

The other manipulation we added was the number of degrees the head would turn, which could either be 0, 15, 30 or 45 degrees (Figure 2f). The 0-degree turn was only used for sequence D, which would be a 180-degree turn. Therefore, we created a total of 16 videos[2], which were evaluated in a within-subjects study.

Videos of each conditions were recorded using Unity. Each video had a length between 10-12 seconds, and featured two people following the robot. After two seconds, the robot would start turning. We added a series of static objects to the scene, so as to give participants the idea of a moving robot.

3.2 Procedure

Participants were recruited through the CrowdFlower platform, and invited to complete a SurveyMonkey questionnaire. Participation was limited to people from European countries. The questionnaire consisted of general information on the experiment and procedure. Following that, participants were shown the 16 videos in a randomized order. For each video participants were asked to indicate on a 7-point semantic differential scale how natural they believed the head gesture of the robot was. Six control questions were added, asking for specific details such as the colors of the various objects, and the turning behavior of the robot. Demographic questions such as gender, age, and nationality were included at the end of the questionnaire, which in total consisted of 27 questions. After completing the questionnaire participants were paid €0.4 through the CrowdFlower platform.

39 participants who failed to correctly answer at least 83.3% of the test questions were excluded from the sample.

3.3 Sample

A total of 29 participants completed the survey, and took on average 8 minutes to complete the survey. The sample consisted of 19 males and 10 females, with an average age of 38.1 (sd=11.9). Participants originated from Germany (34.5%), the United Kingdom (17.2%), Austria (17.2%) and the Netherlands (13.8%).

[1]http://www.crowdflower.com

[2]https://www.youtube.com/playlist?list=PLXeBNyrOHCUG54cewvtaBgR57rUSv1ZNQ

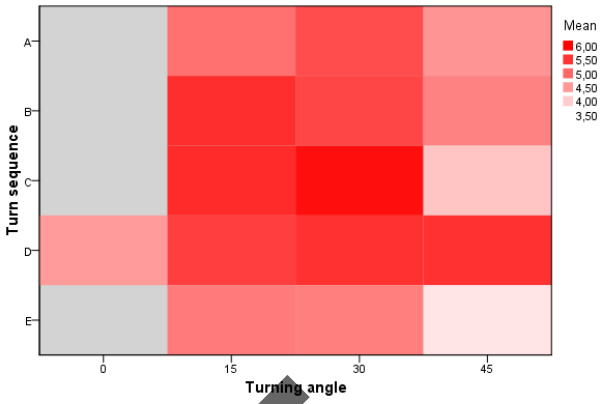


Figure 3: Heatmap

3.4 Results

Mean values for all videos are reported in Table 1 and visualized in Figure 3. As can be seen from Figure 3, there is a preference for turn sequence C at 30 degrees.

Based upon this rapid evaluation we propose to use sequence C30 for the SPENCER robot, which could be implemented in the following way, as also visualized in Figure 2.

4. METHOD

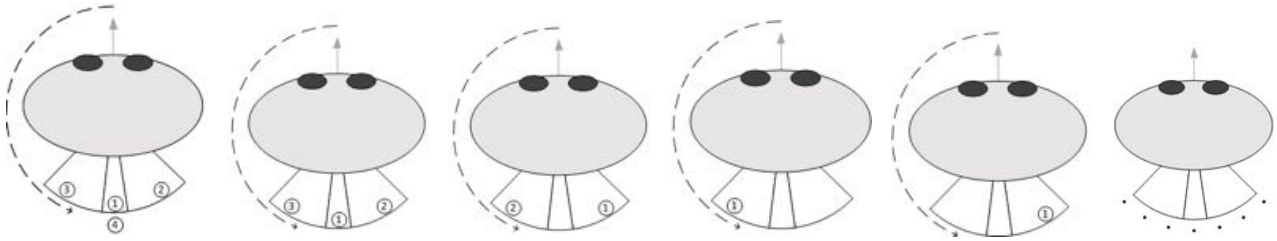
In the previous section we conducted a pilot study, to investigate which head turn was perceived as most natural out of a total of 16 head turns. With the current study we will investigate whether people evaluate the social competence of a guide robot higher when the robot acts with social normative behavior. We have conducted a between-subjects experiment, where we recruited 29 small groups of people who followed a guide robot, and afterwards completed an post-experiment questionnaire.

4.1 Manipulation and apparatus

We manipulated the head behavior of the SPENCER robot to be either socially *normative* or *non-normative*. Based upon our pilot study we implemented a *normative* robot behavior sequence of the head, which consisted of the robot

Table 1: Mean values for each video show that sequence C30 was considered the most normal behavior

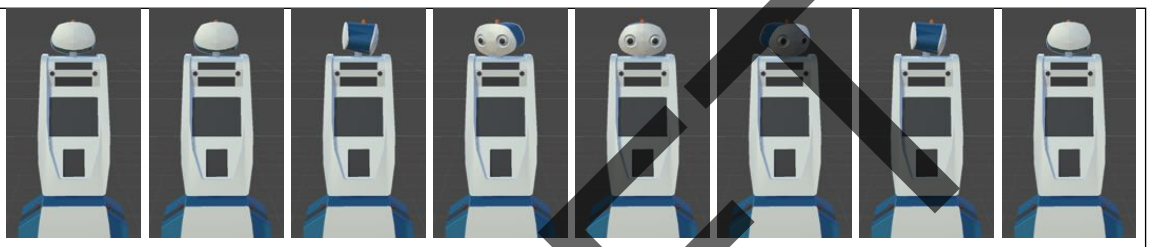
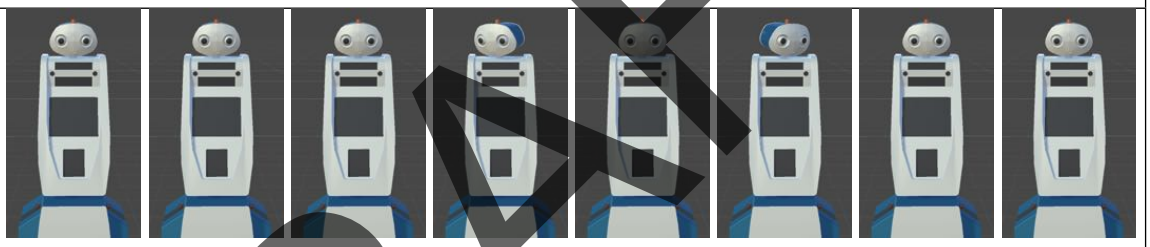
Sequence	Head rotation angle mean score (sd)			
	0°	15°	30°	45°
A		4.9 (2.09)	5.24 (2.44)	4.55 (2.23)
B		5.55 (1.86)	5.31 (2.09)	4.72 (2.10)
C		5.59 (1.92)	5.86 (1.81)	4.07 (2.09)
D	4.48 (1.99)	5.38 (1.82)	5.52 (2.01)	5.52 (2.34)
E		4.79 (1.61)	4.76 (1.86)	3.76 (1.94)



(a) (b) (c) (d) (e) (f)

Figure 2: Five turning sequences were designed for the study (Figures 2a-2e), each head turn sequence was implemented using different turning arcs (Figure 2f).

Table 2: Two 25-second head rotation sequences were programmed on the robot

Normative condition								
Non-normative condition								
Time	0 sec	16 sec	18 sec	19 sec	21 sec	23 sec	24 sec	25 sec

facing forwards, and turning it's head backwards every 25 seconds. In the *non-normative* condition, the head faced the participants, but turned left- and right every 25 seconds, just like the normative condition.

The SPENCER robot is an autonomous guide robot with a 2-DoF turning head, specifically designed to provide services to transfer passengers at international airports [21]. Additional interaction capabilities are provided through it's touchscreen and boarding card reader; these capabilities were not used in this experiment. SPENCER uses four RGD-B cameras and two SICK LMS 500 laser scanners for navigation and obstacle avoidance. In this experiment, the robot drove autonomously to five way-points.

4.2 Task and procedure

The experiment was conducted in a public hallway of an education building. Recruited participants were informed of the goal of the SPENCER robot, and that we were in the final phase of product development. We further told participants that we were interested in their input and feedback on some of the behaviors of the robot. To do so, participants followed the robot while it drove two laps around the hallway and fill out a post-experiment questionnaire afterwards. Participants followed the robot for about 4 minutes (M=273 seconds, SD=52 sec.), see also Table 3).

Table 3: Average task- and robot performance metrics for each trial

	Mean	Sd
Distance	178.33 m	2.275 m
Time to completion	273 sec.	52 sec.
Speed	0.67 m/s	0.095 m/s

After having signed a consent form, the experimenter confirmed the experiment procedure once more. Following this check, the robot started navigating through the hallway (Figure 4). Having followed the robot through the hallway, participants were asked to complete a pen-and-paper based questionnaire consisting of both closed and open questions (Section 4.3).

After having completed the questionnaire all participants were debriefed, in which any remaining questions of the participants were answered. The total duration of the experiment was about 15 minutes. Participants were paid €5 for participating in the experiment.

4.3 Measures

Both objective and subjective measures were collected during the experiment.We annotated the videos collected through the rear-facing RGD-B cameras (Figure 5, Section 4.3.1) and we collected participants' subjective feedback through a questionnaire (Section 4.3.2).

4.3.1 *Video data*

Video data of all experiment trials were collected through 2 cameras mounted on tripods, each recording a section of the hallway (Figure 4), and 2 rear-mounted RGD-B cameras on the robot, recording (part of) the group (Figure 5). Given that these data were recorded from a moving platform, not all participants were always visible. We describe our data analysis method for this particular measure in more detail in Section 4.5.

4.3.2 *Questionnaire*

A 50-item questionnaire consisting of both open and closed questions was designed to measure participants' user experience. Trust in the robot was measured through the competence (6 items, α=0.658) and goodwill (6 items, α=0.692) constructs of the Source Credibility Scale [15]. The perceived anthropomorphism (5 items, α=0.583) and likeability (5 items, α=0.811) scales were used to assess [....] [3].

Attitude towards robots was measured through the NARS [18] interaction with robots (6 items, α=0.657) and social influence of robots (4 items, α=0.675)[3] subscales. Both experience with robots and engagement were measured through one question, and satisfaction with the robot was measured through 3 7-point Likert scaled items (α=0.795) based on [14]. Three open questions inquired whether participants had any further comments and / or suggestions for improvement.

Eight demographic questions provided us with information on participants gender, age, study program and flying experience. Finally, four open questions assessed:

1. What did you like about the robot's behavior?

2. Do you have any comments and/or suggestions regarding the behavior of the robot if it were to guide you at an airport?

3. Would you consider the robot, as you saw it right now, to be a socially aware robot?

4. Which aspect of the guiding do you think should be improved before deploying this robot at an airport?

[3]Item "I feel that in the future, society will be dominated by robots" removed

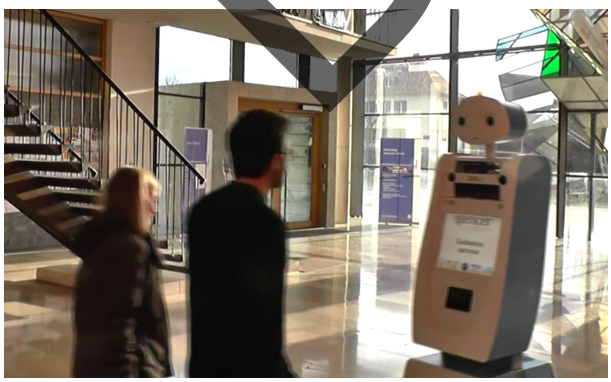


Figure 4: 29 groups of 2-3 persons followed the SPENCER robot

Table 4: 68 participants, divided over 29 groups (between brackets) participated in a between-groups experiment

Condition	2p. group	3p. group	Total
Normative	22 (11)	15 (5)	37 (16)
Non-normative	16 (8)	15 (5)	31 (13)

4.4 Participants

Small groups of 2 and 3 people were recruited in a public building of the Computer Science faculty of the University of Freiburg. A total of 72 participants participated in the experiment, however, due to technical issues with the robot two groups had to be removed from the sample. The final sample consisted of 68 participants, divided over 29 groups (Table 4).

The participants had a mean age of 24.56 (SD=3.861); 73.5% of the sample was male, 26.5% female. Most participants had the German nationality (64.7%), followed by Russian (5.9%), Chinese (4.4%) and Indian (4.4%) participants. 48.5% of the participants had a background in Computer Science, other participants generally had a technical background as well, such as from the field of Embedded Systems Engineering (14.7%) and Microsystems Engineering (16.2%).

Participants' experience with robots was high: 23.5% indicated they had seen robots before, 19.1% of the participants previously worked with robots, and 50% indicated they previously built or programmed robots themselves.

4.5 Data analysis

Internal reliability for all subjective scales were acceptable (see Section 4.3.2). Due to the questionnaire being administered on paper the sample contained 12 missing values. In order to carry-out list-wise comparisons we have replaced these with the means of the respective variable for the purpose of this report.

In order to test for normality a Kolmogorov-Smirnov tests was conducted. All dependent variables were normally dis-

Figure 5: Video data captured through the aft-facing RGB-D cameras

tributed, with the exception of the NARS Interaction subscale. Therefore, in general, we will conduct t-tests and ANOVA's to test our hypotheses. For the NARS Interaction subscale we will conduct Kruskall-Wallis and Mann-Whitney tests instead.

We categorized all answers to the open questions into categories. Some participants chose not to answers one or multiple open questions, while others provided multiple answers. Therefore we use percentages to show differences between the two conditions (Tables 5 - 8).

Engagement was coded by two coders (17% of the data). For each participants we coded when he/she was visible, and whether or not he/she seemed to pay attention to the robot. We finally calculated the users' engagement as a percentage of the time he/she was visible.

Data from the SICK laser scanners was used to calculate mean distance between the participants and robot. During the experiment we observed that participants did not always understand the intention of the robot when turning, therefore we only analyzed the distances at the straight paths (Figure 6).

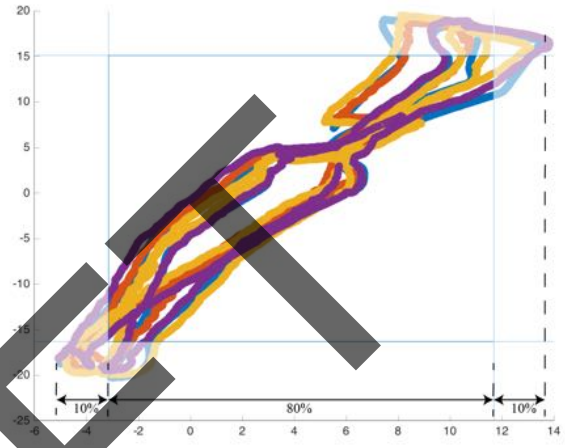


Figure 6: Tracks showing positions of a robot (blue track) guiding a 3-person groups. Highlighted area shows area taken into consideration when calculating features (such as distance)

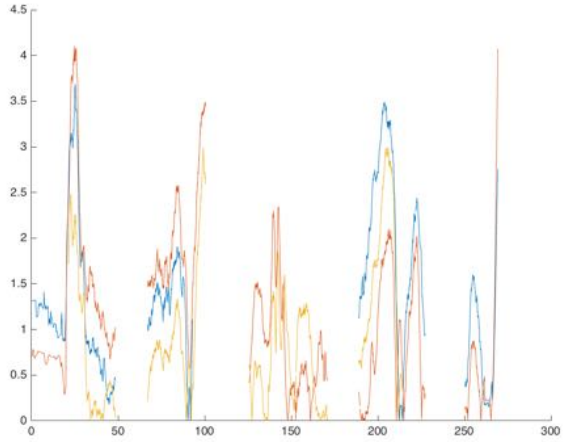


Figure 7: Distance between the robot and 3 participants in the same group. Note the highlighted areas of Figure 6 are not plotted in this graph

5. RESULTS

In this section we will discuss report the results for each of the three hypotheses.

5.1 H1: The operationalization of the socially normative behavior will be evaluated differently

A series of t-tests showed that there was no significant difference between the two conditions on any of the dependent variables, with the exception of the anthropomorphism scale ($t(62)$=-3.867, F=.658, p=0.000); the robot in the non-normative condition was perceived as more anthropomorphic.

Participants were not considered to be significantly more or less engaged in the normative condition (M=78.5%, SD=13.47% compared with the non-normative condition (M=82.0%, SD=14.98%), $T(66)$=-1.029, p=.307.

In the non-normative condition participants kept more distance from the robot (Mdn=1.94m., SD=0.66m.) compared to participants in the normative condition (Mdn=1.81m., SD=0.83m.), $t(66)$=-0.717, p=0.476. During the post-experiment analysis we considered a potential effect of group size on the distance, and though dyadic groups walked closer to the robot (Mdn=1.76m., SD=0.77m.) than triadic groups (Mdn=2.01m., SD=0.71m.) this effect was also non-significant, $t(66)$=0.397 p=0.183.

From the answers of the first question; *"What did you like about the robot's behavior?"*, we gather that participants in particular liked the obstacle avoidance capabilities and the speed of the SPENCER robot (Table 5). In the *normative* condition participants showed higher appreciation for the head turning behavior as compared with the *non-normative* condition, which provides **partial support for H1**.

5.2 H2: Evaluation of the robot's social competence is dependent on attitude towards robots in general

We found significant negative correlations of "negative attitudes towards interaction with robots" on perceived competence (r=-.303, p<0.05), and on goodwill (r=-.261, p<0.05).

Correlations between "negative attitudes towards social influence of robots" were non-significant.

Therefore, we **partially accept H2**, which stated that the evaluation of the robot's social competence is not dependent on socially normative behavior, rather there is a correlation with attitude towards robots in general.

5.3 H3: Evaluation of the robot's social competence is dependent on the anthropomorphic design of the robot

We found significant positive correlations of anthropomorphism on perceived competence (r=.258, p<0.05), and on goodwill (r=.468, p<0.001). Additionally, we found significant positive correlations of likeability on perceived competence (r=.482, p<0.001), and on goodwill (r=.607, p<0.001).

This provides **strong support for H3**, which stated that the evaluation of the robot's social competence is not dependent on socially normative behavior, rather there is a correlation with the evaluation of the anthropomorphic design of the robot.

5.4 Open questions

When asked towards improvements for the SPENCER robot (Table 6), participants indicated would have liked to receive more feedback, for example route information through a map and the progress of the tour. These two categories, making up 43% of the comments have been implemented on the user interface of the robot, but weren't used in the experiment.

Answers to the third questions mainly consisted of short answers; these being "yes", "no" and limited to short sentence such as *"He was aware of who he has to guide, so yes"*. We therefore categorized these answers, as shown in Table 7. It seems that the SPENCER robot was seen as being "socially normal" in general, and we did not see any differences here between the two conditions. However, there were a relative large number of participants who did not provide an answer to this questions (11.76%, 8 participants), therefore it could be that the question was too difficult (e.g. socially normative being a subjective and potentially fuzzy concept).

The final open questions inquired towards capabilities of the SPENCER robot which participants deemed should be improved before being deployed in an actual airport environment. Answers were quite spread out, as can be seen in Table 8. Most participants suggested speed adaptation should be implemented, followed by more feedback to the passengers while they're being guided, including a map and/or time to arrival.

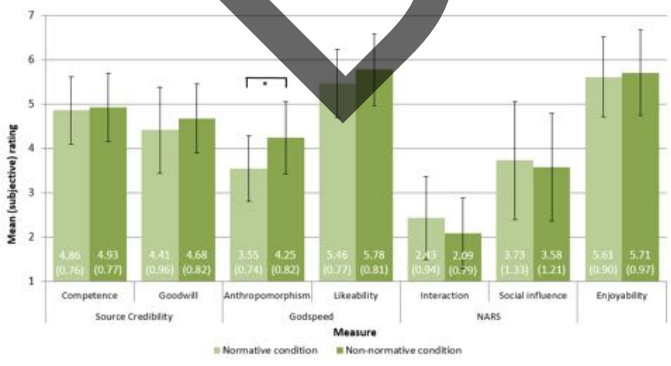


Figure 8: There was no significant difference between the two conditions on any of the dependent variables, * indicate significant difference at p<0.01.

Table 5: Participants in particular liked the obstacle avoidance capabilities of SPENCER; in the *normative* condition participants showed higher appreciation for the head turning behavior.

Category	**Non-normative**	**Normative**	**Total**
(no comment)	2 (6.45%)	3 (8.11%)	5 (7.35%)
Obstacle avoidance	9 (29.03%)	13 (35.14%)	22 (32.35%)
Head appearance	5 (16.13%)	7 (18.92%)	12 (17.65%)
Head checking behavior	4 (12.9%)	14 (37.84%)	18 (26.47%)
Speed	8 (25.81%)	9 (24.32%)	17 (25%)
Robot appearance	1 (3.23%)	1 (2.7%)	2 (2.94%)
(Smooth) movement	2 (6.45%)	5 (13.51%)	7 (10.29%)
Get the job done	3 (9.68%)	0 (0%)	3 (4.41%)
Quiet mechatronics / computers	3 (9.68%)	1 (2.7%)	4 (5.88%)
Total comments	**50**	**35**	**85**

6. DISCUSSION & CONCLUSION

In this paper we presented a study in which we asked small groups of people to follow the SPENCER robot in order to receive feedback on the behavior of the robot in an actual guiding situation. We explored two different behaviors, normative and non-normative behavior, which we operationalized through the head of the robot.

Participants' general subjective assessment of the robot was rather positive: mean ratings for competence (M=4.90, SD=0.762), goodwill (M=4.53, SD=0.888), and likeability (M=5.61, SD=0.798) were the average of the 7-point Likert scale. As expected, the negative attitude towards robot's subscales interaction (M=2.27, SD=0.885) and social influence (M=3.66, SD=1.271) received relative low scores. Self-reported engagement was also high (M=5.66, SD=0.925). This leads us to conclude that the SPENCER robot is a likeable robot which conducts guiding tasks in an effective way. This conclusion is supported by the open questions, which clearly show that participants appreciated the autonomous navigation, including the obstacle avoidance module (Table 5).

Despite our expectations the manipulation with the head direction did not necessarily lead to strong differences on the various items of the questionnaire, contrary our previous pilot study [11]. The open questions did reveal that the head checking behavior in the normative condition was appreciated by participants, though we can conclude that this is not a necessary features for guiding. Therefore, the head could be used for example only to convey the future driving direction of the robot.

Improvements for the SPENCER robot are mostly related to feedback and intention communication; features we delib-

Table 6: As improvement SPENCER could have provided more feedback, for example route information through a map.

Category	Non-normative	Normative	Total
(no comment)	4 (12.9%)	3 (8.1%)	7 (10.3%)
More acceleration/deceleration	2 (6.45%)	2 (5.41%)	(5.88%)
Communicate (motion) intent	2 (6.45%)	2 (5.41%)	(5.88%)
Faster turning arcs	3 (9.68%)	1 (2.7%)	(5.88%)
Feedback: map	5 (16.13%)	9 (24.32%)	(20.59%)
Feedback: progress	8 (25.81%)	7 (18.92%)	(22.06%)
Feedback: auditory	3 (9.68%)	7 (18.92%)	(14.71%)
More fluent movement	2 (6.45%)	2 (5.41%)	(5.88%)
More interaction	4 (12.9%)	7 (18.92%)	(16.18%)
Smooth path planning	1 (3.23%)	0 (0%)	(1.47%)
Speed adaptation	2 (6.45%)	3 (8.11%)	(7.35%)
Total comments	**32**	**40**	**72**

Table 7: 50% of the participants thought the robot was socially normative, 30% thought the robot was not socially normative.

Category	Non-normative	Normative	Total
Unknown / no answer	7 (22.58%)	1 (2.7%)	(11.76%)
No	8 (25.81%)	13 (35.14%)	(30.88%)
A bit / no	0 (0%)	1 (2.7%)	(1.47%)
Neutral	0 (0%)	4 (10.81%)	(5.88%)
More or less	1 (3.23%)	6 (16.22%)	(10.29%)
Yes	15 (48.39%)	12 (32.43%)	(39.71%)
Total	**31**	**37**	**68**

erately not incorporated in the current experiment, but have been tested at the Schiphol Airport during a previous integration event.

To conclude: we conducted a between-groups experiment where a robot guided small groups of 2-3 people in a semi-public space, using one of two different behaviors - socially normative, or non-normative. In general participants were positive about the appearance and behavior of the SPENCER robot, though we did not find differences between the two conditions.

7. ACKNOWLEDGMENTS

This research has been partly supported by the European Commission under contract number FP7-ICT-600877 (SPENCER).

8. REFERENCES

[1] M. Argyle and J. Dean. Eye-contact, distance and affiliation. *Sociometry*, pages 289–304, 1965.

Table 8: Most participants suggested speed adaptation should be implemented, followed by more feedback to the passengers while they're being guided, including a map and/or time to arrival.

Category	Non-normative	Normative	Total
(no comment)	11 (33.33%)	11 (26.83%)	22 (29.73%)
Speed adaptation	4 (12.12%)	11 (26.83%)	15 (20.27%)
Feedback route / flight info	5 (15.15%)	5 (12.2%)	10 (13.51%)
Feedback intention	3 (9.09%)	1 (2.44%)	4 (5.41%)
Feedback shops	1 (3.03%)	2 (4.88%)	3 (4.05%)
Engage conversation.	1 (3.03%)	3 (7.32%)	4 (5.41%)
Carry luggage	1 (3.03%)	2 (4.88%)	3 (4.05%)
Misc	4 (12.12%)	6 (14.63%)	10 (13.51%)
Feedback - Auditory	3 (9.09%)	0 (0%)	3 (4.05%)
Total	**33**	**41**	**74**

[2] C. Bartneck and J. Forlizzi. A design-centred framework for social human-robot interaction. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, pages 31–33, 2004.

[3] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[4] J. M. Beer, A. Prakash, T. L. Mitzner, and W. A. Rogers. Understanding robot acceptance. 2011.

[5] C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1):119–155, 2003.

[6] B. R. Duffy. Anthropomorphism and the social robot. *Robotics and autonomous systems*, 42(3):177–190, 2003.

[7] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[8] M. M. A. Graaf. *Living with robots: investigating the user acceptance of social robots in domestic environments*, volume 15. Universiteit Twente, 2015.

[9] M. Heerink et al. Assessing acceptance of assistive social robots by aging adults. 2010.

[10] M. A. Hollands, A. E. Patla, and J. N. Vickers. "look where you're going!": gaze behaviour associated with maintaining and changing the direction of locomotion. *Experimental brain research*, 143(2):221–230, 2002.

[11] M. Joosse, R. Knuppe, G. Pingen, R. Varkevisser, J. Vukoja, M. Lohse, and V. Evers. Robots guiding small groups: the effect of appearance change on the user experience. In *Proceedings of the 4th Internatioal Symposium on New Frontiers in Human-Robot Interaction*. University of Kent, 2015.

[12] D. E. Karreman, G. U. S. Bradford, E. M. van Dijk, M. Lohse, and V. Evers. Picking favorites: The influence of robot eye-gaze on interactions with

multiple users. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 123–128. IEEE, 2013.

[13] D. E. Karreman, G. D. Ludden, E. M. van Dijk, and V. Evers. How can a tour guide robot's orientation influence visitors' orientation and formations? In *Proceedings of the 4th Internatioal Symposium on New Frontiers in Human-Robot Interaction*, 2015.

[14] K. M. Lee, W. Peng, S.-A. Jin, and C. Yan. Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of communication*, 56(4):754–772, 2006.

[15] J. C. McCroskey and J. J. Teven. Goodwill: A reexamination of the construct and its measurement. *Communications Monographs*, 66(1):90–103, 1999.

[16] Merrian-Webster. Communication, July 2015.

[17] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley. *Robotics & Automation Magazine, IEEE*, 19(2):98–100, 2012.

[18] T. Nomura, T. Kanda, T. Suzuki, and K. Kato. Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *Robotics, IEEE Transactions on*, 24(2):442–451, 2008.

[19] S. Norris. *Analyzing multimodal interaction: A methodological framework*. Routledge, 2004.

[20] M. Shiomi, T. Kanda, H. Ishiguro, and N. Hagita. A larger audience, please!: encouraging people to listen to a guide robot. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 31–38. IEEE Press, 2010.

[21] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, et al. Spencer: A socially aware service robot for passenger guidance and help in busy airports. 2015.

[22] M. von Grünau and C. Anston. The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, 24(11):1297–1313, 1995.

Appendix B: Feedback questionnaire

Feedback questionnaire SPENCER robot

Thank you for your interest in the SPENCER robot. SPENCER is a KLM project is collaboration with the European Commission. The goal is to develop a robot that can guide passengers from A to B. With this survey we'd like to find out how this robot is being perceived, and which aspects of the robot could be improved. Thank you in advance for completing this questionnaire.

At which airport did you start your journey?	
Which airport are you travelling to?	
How many times have you flown in the last 12 months?	
Including yourself, how many people are travelling in your immediate group?	
What is the chief purpose of your present trip?	O Business O Pleasure

Please indicate your level of agreement:	Not at all			Neutral			Very much so	Undecided
Do you have a general interest in technology?	O	O	O	O	O	O	O	O
Do you have an interest in robots?	O	O	O	O	O	O	O	O
Was it easy to interact with Spencer?	O	O	O	O	O	O	O	O
How satisfied were you with the information SPENCER provided?	O	O	O	O	O	O	O	O
Did SPENCER react fast enough?	O	O	O	O	O	O	O	O
Did SPENCER drive fast enough?	O	O	O	O	O	O	O	O
Did you enjoy your experience with SPENCER?	O	O	O	O	O	O	O	O
Did you understand where SPENCER was going?	O	O	O	O	O	O	O	O
Did you receive enough feedback from SPENCER?	O	O	O	O	O	O	O	O
Do you think SPENCER acted appropriately?	O	O	O	O	O	O	O	O
Do you think SPENCER is friendly?	O	O	O	O	O	O	O	O
Do you think the way the robot behaved was appropriate?	O	O	O	O	O	O	O	O
Would you trust Spencer to guide you next time you have to transfer at an airport?	O	O	O	O	O	O	O	O
Would this improve your satisfaction with KLM	O	O	O	O	O	O	O	O

What is your gender?	O Male O Female
What is your age?	

If you have any further comments, please write them in this box (or at the back of this form)

Thank you for completing this questionnaire and providing us with your input for our work.

spencer Social situation-aware perception and action for cognitive robots

SEVENTH FRAMEWORK PROGRAMME

KLM Royal Dutch Airlines

Feedback questionnaire SPENCER robot

Thank you for your interest in the SPENCER robot. SPENCER is a KLM project is collaboration with the European Commission. The goal is to develop a KLM robot that offers assistance in the airport. With this survey we'd like to find out how you experienced the robot, and which aspects of the robot could be improved. Thank you in advance for completing this questionnaire.

At which airport did you start your journey today?	
Which airport are you travelling to?	
Including yourself, how many people are travelling in your immediate group?	
Is your trip for business or pleasure?	O Business O Pleasure

To what extend do you agree with the following statements:	Not at all			Neutral			Very much so
I think the robot is very likeable	O	O	O	O	O	O	O
I think the robot has an appealing design	O	O	O	O	O	O	O
I think the robot is intelligent	O	O	O	O	O	O	O
It was easy to understand the robot's intentions	O	O	O	O	O	O	O
My experience with the robot was very positive	O	O	O	O	O	O	O
The robot is unreliable	O	O	O	O	O	O	O
The robot acted appropriately	O	O	O	O	O	O	O
This service improves my customer satisfaction	O	O	O	O	O	O	O
The robot functioned well	O	O	O	O	O	O	O
It was difficult to use the robot	O	O	O	O	O	O	O

	Very negative			Neutral			Very positive
What is your opinion of robots in general?	O	O	O	O	O	O	O

What is your gender?	O Male O Female
What is your age?	

Having followed this robot, would you consent to us using the (video)data collected in scientific presentations to other scientists? If so, please sign in the box to the right.	

Thank you for completing this questionnaire and providing us with your input for our work.