D2.1 – Localization part 1

Project acronym: *TERESA*

Project full title: Telepresence Reinforcement-Learning Social Agent

Grant agreement no.: 611153

Due-Date:	15
Delivery:	Month
Lead Partner:	University of Twente (UT)
Dissemination Level:	PU
Status:	Submitted
Version:	1.0

DOCUMENT INFO

Date and version number	Author	Comments
10.02.2014 v0.1	Jered Vroon	First outline.
19.02.2014 v0.2	Jered Vroon	Updated outline, wrote introduction, aims and methods/set-up
23.02.2014	Jie Shen, Stavros Petridis	Provided texts describing the placement of the Dalsa camera, the localization based on face detection and the evaluation thereof.
25.02.2014 v0.3	Jered Vroon, Gwenn Englebienne	Added description of the data set, evaluation of localization algorithms and conclusions.
27.02.2014 v1.0	Jered Vroon, Gwenn Englebienne, Vanessa Evers	Final version.

TABLE OF CONTENTS

1 Executive summary 5

2 Contributors 6

3 Localization 7

3.1 Aims 7

3.2 Sensor set-up 8

3.2.1 Kinect placement and recording 8

3.2.2 Dalsa placement and recording 10

3.3 Methods 10

3.3.1 Acquiring a ground truth 12

3.4 Data set 12

3.4.1 Upper bound on persons within Kinect sensor range 13

3.4.2 Occurrence of occlusions 14

3.5 Evaluation of localization algorithms 14

3.5.1 Performance of the Kinect skeleton tracker 15

3.5.2 Performance of face-detection based localization 17

4 Conclusions 19

5 References 21

LIST OF TABLES AND FIGURES

Figure 1 - Teresa robot with the mounted sensors ... 9
Figure 2 - Dalsa placement (original plan) ... 10
Figure 3 - Dalsa placement (revised plan) ... 10
Figure 5 - Example of the interactions in the collected data. A person controlling a Teresa robot approaches, converses with and retreats from a group several times. ... 11
Figure 5 - Head and back markers used for acquiring ground truth ... 11
Figure 6 - Example of tracking a face from the data collected ... 18

Table 1 - Theoretical limits on the detection capabilities of the Kinect sensor, in percentage of the frames recorded. ... 13
Table 2 - Theoretical upper bound (UB) on the detection capabilities of the Kinect sensor compared with the actual detections of the Kinect skeletal tracker (LB). Both in percentage of the frames recorded. The last row shows the ratios of the LB to the UB in percentages. ... 16
Table 3 - Theoretical upper bound (UB) on the detection capabilities of the Kinect sensor compared with the actual detections of the Kinect skeletal tracker (LB) for one of the groups. Both in percentage of the frames recorded. The bottom row shows what would have been the upper bound if all missing frames had instead been interpreted as having 0 people within sensor range. ... 16

1 Executive summary

This deliverable, together with its next iteration, aims to introduce a complete localization component for multiple people detection for the Teresa project. This localization is based on data from a depth sensor and will also integrate data from face detection.

In the beginning of the project we have, in collaboration with the project partners, devised a suitable way to mount a Kinect depth sensor and a Dalsa camera for face detection on the Giraff robot (see D6.2 for a full description).

In November 2014, we have collected a large data set at the UT. It contains both data from this Kinect sensor mounted on the robot and sub-centimetre precision tracking data which can be used as a ground truth. Recordings involve a wide range of natural interactions, including various approaches and retreats. This data set is, particularly because of the availability of a ground truth, suitable for the training and testing of various algorithms for automatic pose estimation and analysis of people in the robot's vicinity.

We have cleaned this data set and used it to investigate if the current sensor mounting is suitable for the needs of the project. In addition, we have used it to show that (partial) occlusions are relatively uncommon (3.2% of the frames). We have also applied an existing algorithm that can do localization, the Kinect skeletal tracker, to the data to investigate how well it is able to detect and localize people from the data. Compared to the upper bound of people being in sensor range, the skeletal tracker correctly detects at least one person in at least 68.7% of the frames.

We additionally discuss the algorithms used for tracking people based on face detection and its current limitations.

As proposed in the "Request for Amendment No. 1 to grant agreement No. 611153 - Project title 'Teresa'", of February 3rd 2015, submitted by project coordinator Shimon Whiteson, this deliverable is the first iteration of a recurring deliverable. This will allow the UT and ICL to put more work carried out for T2.2 (which runs to M33) in D2.1. This is particularly relevant since Ronald Poppe left the UT in May 2014 and has been replaced only from January 2015 on, when Gwenn Englebienne started at the UT as an assistant professor working on the project – which has limited the PM's invested in the task during the first year.

In this first iteration we have investigated a system that can do basic localization based on the Kinect data and face detection. The next iteration (M15+12) will update this functionality and add automatic detection of those body pose features that were found to be particularly relevant for groups in T3.1. We will also investigate ways in which multiple people can be tracked based on the face detection. Lastly, we will integrate the localization based on the Kinect data with the localization based on face detection to further improve performance and robustness.

2 Contributors

All efforts related to the depth sensor based localization described in this document are the collaborative effort of Jered Vroon, Gwenn Englebienne and Vanessa Evers (UT). Jered Vroon wrote the first drafts of this document, based on results by him and Gwenn Englebienne. Vanessa Evers and Gwenn Englebienne supervised these efforts and gave feedback on these drafts.

The efforts related to the face detection based localization described in this document are the collaborative effort of Jie Shen and Stavros Petridis (ICL).

Ronald Poppe (UT) initiated the work on this deliverable and played an active role in devising a suitable way to mount the Kinect sensor on the Giraff robot in collaboration with the consortium partners. Jie Shen has been involved in devising and revising a suitable way to mount the Dalsa sensor. Lasse Hedman (GT) has been responsible for developing the required mounts. Noé Perez Higueras and Ignacio Perez Hurtado De Mendoza (UPO) have done most work in setting up the software and hardware to collect data from the Kinect sensor using rosbag.

In addition, we wish to mention here that the collection of data with ground truths here described was conducted as part of a larger, collaborative effort. This effort was the joint work of Jered Vroon, Michiel Joosse, Manja Lohse, Jan Kolkmeier, Jaebok Kim, Khiet Truong, Gwenn Englebienne, Dirk Heylen, and Vanessa Evers (UT)[1]. Jan Kolkmeier created the markers and configured the system we used to acquire the ground truth.

The continued effort on the next iteration of this deliverable will be the joint work of Jered Vroon and Gwenn Englebienne at the UT and Jie Shen and Stavros Petridis at ICL.

[1] We declare that all of the hours invested in this joint effort have only been written once. More specifically, only Jered Vroon, Jaebok Kim and Gwenn Englebienne have written hours invested in this study on the Teresa project. Since the results of the study have been used for tasks in work packages 2 and 3, these hours have been divided accordingly.

3 Localization

The TERESA project aims to develop the modules required for a telepresence robot to display socially appropriate behaviour in various interactions with groups of (and individual) elderly. These interactions entail both navigation in areas with people and conversation behaviour, including the approach and retreat behaviours required to socially initiate or terminate a conversation.

One of the modules that is required for interaction with people is the capacity to detect and localize them. This deliverable, together with its next iteration, is aimed at introducing a complete localization component for multiple people detection for the Teresa project. This localization is based on data from the Kinect depth sensor and can also integrate data from face detection.

As we will describe in more detail in our aims (Section 3.1), in this first iteration we will focus on two existing algorithms for localization based on the used Kinect depth sensor and face detection. These are the Kinect skeletal tracker developed by Microsoft and the face detection algorithms developed by ICL respectively. In this document, we will specify the sensor set-up (Section 3.2), and describe our methodology (Section 3.3) for collecting a data set (Section 3.4) that we used to evaluate the performance of those algorithms (Section 3.5). We will discuss the implications and limitations of our findings in more detail in our conclusions (Section 4). There we will also describe how the next iteration of this deliverable will build on the results from this iteration.

3.1 Aims

In the description of work, this deliverable is described as a complete localization component for multiple people detection. It is related to parts of to Task 2.2, which further specifies the aims to handle occlusions and to do more detailed pose estimation that will help understand the social (group) dynamics.

For this first iteration of the deliverable, we have conducted a range of activities that we will all describe in this report;

- Devise a suitable way to mount and set-up the Kinect and Dalsa such that they can collect the data required to do localization
- Acquire a data set with ground truth
- Use this data set to evaluate and compare the different mountings
- Use this data set to investigate how commonly occlusions occur

- Apply the Kinect skeleton tracker developed by Microsoft to the data set and evaluate its performance in terms of detecting people.
- Evaluate the localization based on the face tracking developed by ICL.

Though there thus already exist suitable algorithms for localization, for the next iteration of this deliverable we aim to improve on these algorithms. Specifically, we aim to;

- Extend on the localization based on the depth sensor by adding features that provide information on the social (group) dynamics
- Improve the localization based on the depth sensor, for example by investigating suitable ways to handle occlusions
- Extend the localization based on the face detection, such that it can detect multiple faces and identify different people by their face.
- Combine the information from the depth sensor and the face detection to improve the localization

This way, the work in the two iterations of this deliverable will together form a complete localization component for multiple people detection that combines information from face detection and a depth sensor with information on social (group) dynamics.

3.2 Sensor set-up

In this section we will describe the set-up of the Kinect sensor (Section 3.2.1) and the Dalsa sensor (Section 3.2.2), both in terms of physical placement on the robot and in terms of software used to collect data. Because these have also been described in Deliverable 6.2, we will here focus on the requirements that follow from the aim of using these sensors for localization.

3.2.1 Kinect placement and recording

The main consideration in placing the Kinect depth sensor was to ensure a proper range in which a group of people positioned relatively close to the Teresa robot could still be detected. Because the Dalsa camera is aimed at collecting face data, we considered getting a full view of the people to be less relevant than getting a partial view of as many people as possible.
For this reason, as described in more detail in D6.2, we decided to mount the Kinect depth sensor horizontally on the Kinect on the base of the Giraff robot (see Figure 1). We tilted the sensor up slightly (12°) to provide a good view centred on the waist of people at about 1-2 meters from the robot – which is expected to be the common minimum distance.

A limitation of the chosen position is that it will probably not work when the robot is close to a table to interact with people seated at that table. However, since the robot will then no longer be navigating, the localization information is less relevant at that stage. We used the Kinect V1 for Xbox 360.

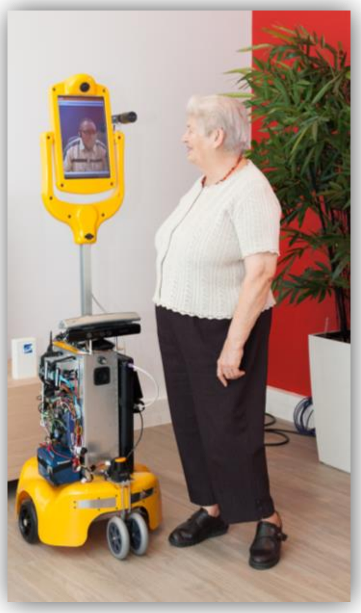

Figure 1 - Teresa robot with the mounted sensors

Two set-ups have been created with which data from the Kinect depth sensor can be recorded. In the first, the Kinect is connected to the laptop running Windows that is mounted on the left of the Teresa robot, which collects data (30fps) using the Kinect studio application v1.7 developed by Microsoft[2]. In the second, the Kinect is connected to the motherboard running ROS on the right of the Teresa robot, which collects data using rosbag[3]. To reduce the required storage space (which is limited on the motherboard), we configured rosbag to store 5fps per second. Because the position of people usually does not change that quickly, we expect this framerate to be sufficient. This second set-up will be used within the architecture of the Teresa project as it is most compatible with the resource allocation of the consortium partners.

In both set-ups, the collected data is 640x840 RGB data, and 320x240 depth data.

[2] https://msdn.microsoft.com/en-us/library/hh855389.aspx

[3] Rosbag is a ROS package for recording and playing back data from various sensors, http://wiki.ros.org/rosbag. Rosbag can also be used with the Kinect sensor, though only with the Kinect for XBOX 360 (not with the Kinect for Windows), instructions can be found at http://wiki.ros.org/openni_launch/Tutorials/BagRecordingPlayback.

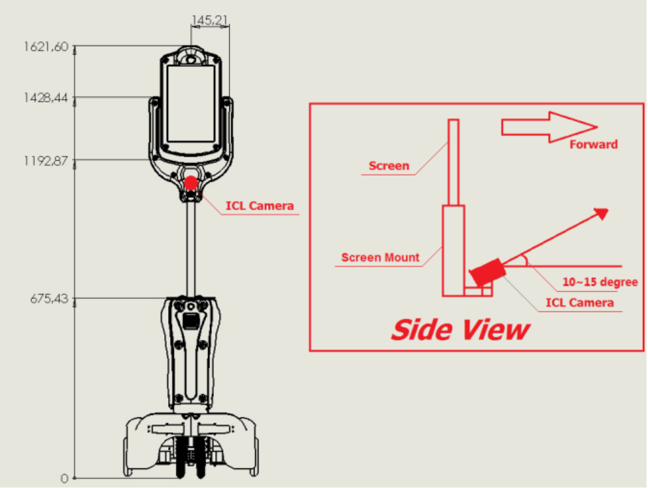

Figure 2 - Dalsa placement (original plan)

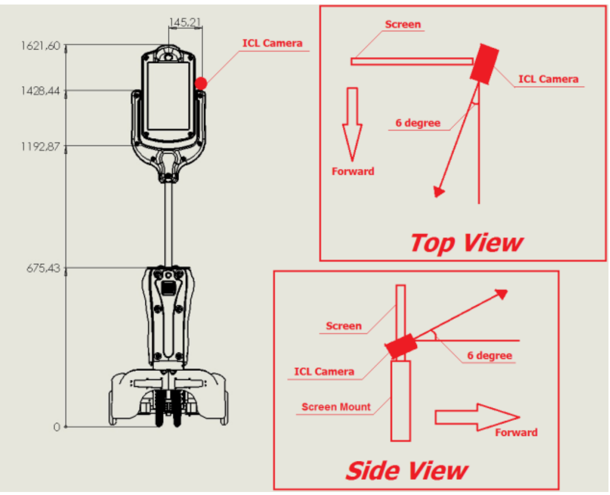

Figure 3 - Dalsa placement (revised plan)

3.2.2 Dalsa placement and recording

ICL's component uses the raw video stream acquired from a Dalsa HM-1400 camera as input. The video stream is captured with a resolution of 1400x1024 pixels, at 64 (on the Teresa robot) or 50 (on the controller-side) frames per second.

When deployed to the Teresa robot, we have tested two plans for camera placement, as illustrated in Figure 2 (the original plan) and Figure 3 (the revised plan). In the original plan, the camera was mounted below the robot's screen and tilted upward at 10-15 degrees. This design has been proven problematic. Even at 15 degrees, from 1 meter away, the camera would be pointing to a height of 1367.9mm, which is still too low to capture the face of average adults. Tilting the camera further upward, however, would significantly degrade our tracker's accuracy because our algorithm was optimised toward near-frontal face images. Therefore, we have revised the plan to mount the camera at a higher location. As shown in Figure 3, the camera is now placed on top of the screen's supporting beam, tilted 6 degrees upward and 6 degrees inward. With this configuration, peoples' faces would be located in the centre of the captured video when they are standing at 1 meter away from the robot, facing toward the screen. Another benefit of the current configuration is that it also allows us to capture faces (generally) closer to frontal pose. We used a C-mount lens on the camera with fixed focus (1 meter) and aperture value (F/4).

3.3 Methods

As discussed in the introduction and aims, we needed a data set with ground truth in a suitable context. This data set will allow us to test the performance of localization algorithms,

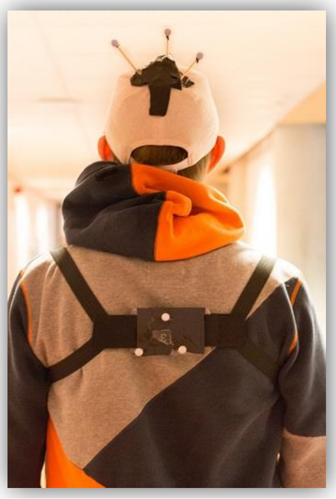

Figure 5 - Head and back markers used for acquiring ground truth

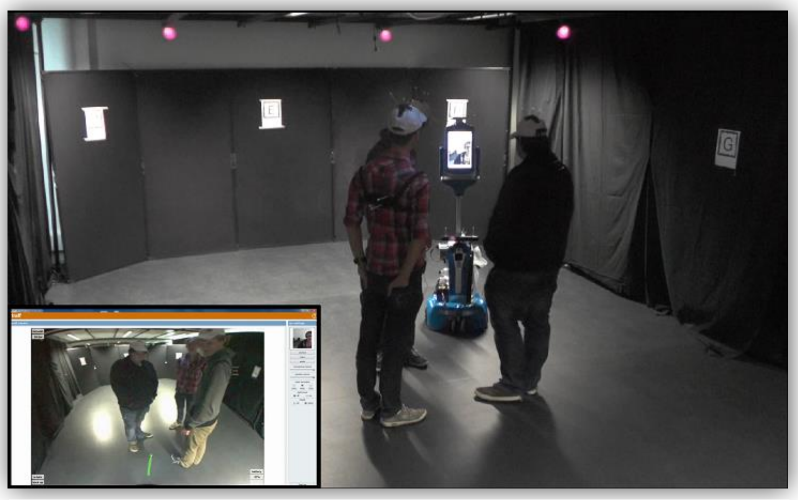

Figure 5 - Example of the interactions in the collected data. A person controlling a Teresa robot approaches, converses with and retreats from a group several times.

to train them, and to evaluate the requirements that the context places on for example the placing of the sensors.

We set up a data collection with this aim. In line with the contexts described in the project, we created a set up in which the Teresa robot was used multiple times to approach small groups (of three people), have a conversation with that group, and then retreat from it (see Figure 5 and Figure 5). The interactions took place in a clean, non-cluttered environment.

For our data collection, we used different groups of 4 participants, one of which controlled the robot, while the others formed the groups with which the robot interacted. In each group of participants, the Teresa robot was used in 8 ‘cycles’ to approach the group members, converse with them and retreat from the group members again. To ensure variety in the data, the direction from which the robot had to approach and to which it had to retreat was different in each of these cycles. We indicated these directions with markers, evenly spaced around the area in which the interaction took place, and varied the order in which these directions were used between the different group of participants. After the 8th cycle, there was one last approach, after which the participants were given some time to wrap up their conversation.

During the interactions, we collected data with the Kinect sensor using Kinect studio, as described in Section 3.2.1. In addition, we also collected ground truth data on the position of the robot and the group members, which we will describe in more detail in Section 3.3.1. We also recorded the interactions using two camera’s (one side view, one fish eye top down view) and by using ScreenCapture software to capture the interactions with the Giraff interface by the person controlling the Giraff.

This data collection was part of the study into socially normative robot behaviour that is described in more detail in Deliverable 3.1.

3.3.1 Acquiring a ground truth

All three group members were equipped with uniquely identifiable markers (one on the back of the chest, one on a cap, see Figure 5), which were tracked by an OptiTrack motion capture system[4] using 8 infrared cameras. This system allows sub-centimetre level precision tracking of both position and orientation of each marker with 120fps. The robot was equipped with one marker on its ‘head’ and one on the Kinect sensor mounted on its base. This last marker gave us the position and orientation of the Kinect in each frame, which allowed us to investigate the data from the perspective of the Kinect.

We optimized tracking for the centre of the interaction area, to make sure we could properly capture the interaction. Markers near the edges of the interaction area could often not be tracked. To ensure proper tracking of the actual interaction, we informed the group members about this and asked them to stay roughly in the centre of the interaction area.

The markers worn by the members of the group could be in a slightly different orientation depending on how they were worn by those group members. For this reason, at the beginning of the data collection we asked all three group members to stand on a line, facing one of the walls in their default upright position. This data was used to calibrate the recordings.

3.4 Data set

A total of 56 participants participated in our data collection, divided into 14 groups of 4 persons. Of these, 13 (23.2%) were female, 43 (76.8%) male. All were students, having an age between 18 and 32 years with a mean of 20 (SD=2.2). Most participants had the Dutch nationality (85.7%). This resulted in data on a total of 126 approaches and 112 conversations and retreats. In total, we collected Kinect data with ground truth covering well over six hours of interaction.

The collected data from the various different sensors was manually synchronized. We found some gaps in the collected Kinect data, usually around moments where the robot made an abrupt stop. We expect that the used hard drive (an HDD rather than an SSD) was not robust to these movements, this will be remedied in future experiments/data collections. Markers near the edges of the area in which the interaction took place were often not tracked.

The data from this data set can be shared with all consortium partners as required and parts of it may in a later stage be made available to the community.

In the remainder of this section, we will describe quantified observations we have made with the data set. We have investigated in the data set how many people will at maximum be

[4] www.naturalpoint.com/optitrack/

within the sensor range of the Kinect sensor in its current and in alternative placements (Section 3.4.1). We have also investigated in the data set how commonly people in a group are (partly) occluded by one of their fellow group members from the perspective of the Kinect sensor (Section 3.4.2).

3.4.1 Upper bound on persons within Kinect sensor range

In the dynamic situation of the interaction, the robot-mounted Kinect's performance is defined by a number of factors:

1) the Kinect's hardware and its detection algorithms, which we evaluate in section 3.5.1;
2) the location and orientation of the robot with respect to the other participants; and
3) the mounting of the Kinect on the robot.

In this section, we evaluate the effect of the latter two aspects on the Kinect's performance. We consider two different ways to mount the Kinect on the robot: vertically and horizontally: the standard horizontal mounting has the advantage that a wider area is observed, at the cost of not seeing the head and feet of people standing close by, while the standard horizontal mounting allows for a better perception of people inside the field of view, at the expense of a more restricted field of view.

Table 1 shows the theoretical detection limits for the Kinect sensor, for the two possible mounting methods. We consider that the Kinect could theoretically detect a person if the centrepoint of the person falls to within 5° of the Kinect's field of view. This corresponds to saying that at least part of the person must be visible.

The first column indicates how frequent the robot was out of the OptiTrack's range; 27.6 percent of the time. Of the remaining frames, the robot and its Kinect were positioned such that the corresponding number of people could be detected. It is important to note how the robot and participants position themselves so that the robot is usually facing at least one person. As a consequence, the vertical or horizontal mounting of the Kinect have little effect on whether or not the Kinect can detect anybody. However, if we look at the possible detection rates for all interaction partners in the group, a very different picture arises and the

	Out of range	0	≥ 1	≥ 2	≥ 3
Vertical placement	27.6	27.2	72.8	35.8	9.7
Horizontal placement	27.6	26.1	73.9	55.6	20.7

Table 1 - Theoretical limits on the detection capabilities of the Kinect sensor, in percentage of the frames recorded.

value of the horizontal placement becomes obvious: overall, the Kinect can track all human participants 20.7% of the time in the horizontal position, as opposed to 9.7% of the time when in vertical position. Given the importance of all members in the interaction, we focus on the horizontal Kinect placement in the remainder of this evaluation.

3.4.2 Occurrence of occlusions

The location of the robot with respect to the other participants affects their perception in a number of ways, most important of which are the field of view of the sensor (discussed above) and the relative position of the participants which can lead to occlusions. In human-only interactions, the participants tend to position themselves so as to allow all participants in the group to have an unhindered view of all others, so that occlusions are virtually non-existent. With the addition of a robot, things are slightly different: not only is its field of view less restricted, the robot is not as nimble as most human participants, making its positioning in the group less seamless and occasionally leading to occlusions.

We therefore evaluated the importance of occlusions in our data by computing how often the robot and two human participants are approximately in a single line, and conclude that occlusions are not an important problem in the interaction. We assume at least partial occlusion when the angle between the lines from the robot to two participants is less than 10° and found that, overall, only 3.2% of the frames contained the partial occlusion of one participant by another from the robot’s perspective. Those occlusions mostly happen during the approach stage, before the robot is actually part of the group, and are therefore not a major problem. The limited field of view of the sensor, which makes it difficult for the robot to perceive three people when part of the group, is a bigger limitation.

3.5 Evaluation of localization algorithms

In this section we will introduce two algorithms for localization and evaluate them. We will evaluate the Kinect skeleton tracker developed by Microsoft against the upper bound introduced in the previous section. We will also introduce tracking based on face detection and discuss its current limitations.

3.5.1 Performance of the Kinect skeleton tracker

We use the Kinect skeletal tracker[5] to detect people in the collected Kinect data. This is the algorithm designed by Microsoft for the purpose, it is optimized to track people directly facing the sensor and can detect up to six people, two of which can be tracked in detail. One limitation of the Kinect skeletal tracker is that it only works on data stored in the format used by Microsoft's Kinect studio (.xed).

Because of the way in which the data was stored, we could not access the data as is, but instead were required to play it back in real-time. Therefore, we wrote a program that uses the Kinect skeletal tracker to do online detection and tracking of people and applied it to such playbacks of the data.

We observed no false positives in the detection by the Kinect skeletal tracker. This is in line with what we would expect based on the data; because we collected the data in a non-cluttered environment, all data-points belonged to either walls or to people[6] – making human-like patterns in the data that did not match to actual people unlikely. Though in rare cases the Kinect skeletal tracker detected four people, this fourth person always turned out to be the experimenter overseeing the interactions who accidentally was within sensor range.

3.5.1.1 Comparison with upper bound

If being within sensor range, as discussed in Section 3.4.1, is considered as an upper bound, the performance of the Kinect skeleton tracker can be considered as a lower bound. Here we will compare these two to each other for the full data set, see Table 2. As can be seen in the table, the ratio for detecting at least one person of the Kinect skeletal tracker to this upper bound is 68.7%.

Making the simplified assumption that the data is comparable, this indicates that the Kinect skeletal tracker can detect at least one person in a reasonable 68.7% of the cases in which at least one person is in sensor range. This assumption is simplified, for example because it does not take into account the fact that there were many more frames missing from the upper bound (27.6%) than from the Kinect skeletal tracker (3.4%). We will investigate this assumption in more detail in Section 3.5.1.2.

The ratios for detecting at least two persons (47.8%) are lower than those for detecting at least one. Those for detecting at least three are even lower (25.6%). This can partly be explained by considering the task of detecting at least two people as equal to the task of detecting one person twice (68.7% x 68.7% ≈ 47.8%). However, the ratio for detecting three

[5] https://msdn.microsoft.com/en-us/library/hh973074.aspx

[6] The only other object in the room was the robot. Since the Kinect sensor was mounted on the robot, the robot was never seen by the sensor.

people is even lower than that reasoning would predict (68.7% x 68.7% x 68.7% = 32.4%), suggesting that detecting three people presents additional difficulties to the skeletal tracker when compared to detecting one or two.

	Missing	0	≥ 1	≥ 2	≥ 3
Kinect detection (LB)	3.4	49.2	50.8	26.6	5.3
Upper bound (UB)	27.6	26.1	73.9	55.6	20.7
LB / UB	123.2%	188.5%	68.7%	47.8%	25.6%

Table 2 - Theoretical upper bound (UB) on the detection capabilities of the Kinect sensor compared with the actual detections of the Kinect skeletal tracker (LB). Both in percentage of the frames recorded. The last row shows the ratios of the LB to the UB in percentages.

3.5.1.2 Detailed comparison for one group

To further investigate the relationship between this upper and lower bound, we additionally performed a more detailed comparison for one group. With this comparison, we aimed both to get more insight into the specific situations that pose a challenge for the Kinect skeletal tracker and to further investigate the simplified assumption that the data is comparable. The comparison between the upper bound and the Kinect skeletal tracker for this group can be found in Table 3.

GROUP 4	Missing	0	≥ 1	≥ 2	≥ 3
Kinect detection (LB)	7.8	39.8	60.2	40.5	2.78
Upper bound (UB)	26.2	24.3	75.7	64.5	17.8
UB, treating missing as 0	0	40.0	60.0	51.1	14.1

Table 3 - Theoretical upper bound (UB) on the detection capabilities of the Kinect sensor compared with the actual detections of the Kinect skeletal tracker (LB) for one of the groups. Both in percentage of the frames recorded. The bottom row shows what would have been the upper bound if all missing frames had instead been interpreted as having 0 people within sensor range.

Kinect skeletal tracker: By looking at the data of this group, we observed that the Kinect skeletal tracker needs to “catch on”. This takes some time and most commonly happens in frames where the Kinect data provides a full frontal, largely non-occluded view of a person. It should be noted that in our data, we often saw that the group members actively turned to the robot and tried not to stand behind other people, causing these non-occluded frontal views to naturally arise.

The Kinect seems to perform below the upper bound mostly if there was no opportunity to “catch on”, for example due to a limited field of view, occlusions or side views. This could also explain why detecting a third person is hard for the skeletal tracker. Likewise, we observed that when approaching a group the robot was at a distance from them, causing more occlusions and more side views; during approaches the Kinect skeletal tracker usually performed below the upper bound.

We additionally observed that, after it catches onto someone, the Kinect skeletal tracker it manages very well to keep track of that person. In some cases it could even exceed the theoretical upper bound by tracking people by just their extremities.

Comparable data: The upper bound and the skeletal tracker appeared to be quite comparable, with two notable exceptions.

First, as stated before, it happened that the Kinect skeleton tracker also tracked the experimenter, who was not tracked by the OptiTrack. This will have favoured the lower bound in the comparison. However, it should be noted again that this was very rare, as the protocol was designed to avoid this situation as much as possible.

Second, we observed a large discrepancy in the missing frames between the upper and the lower bound. When there was no OptiTrack tracking available, this usually corresponded with (the end of the) retreat movement, when the robot was facing the wall and had no people within sensor range. In rare cases it corresponded with the approach, but then usually only very shortly, during the beginning of the approach. It was very rare during conversation, and then usually only lasted a few frames (and was caused by occlusions of the trackers on the robot for the OptiTrack cameras by the group members). At the same time, missing Kinect frames usually corresponded with the motions at the end of the approach, when the robot was usually facing the group. This discrepancy will have favoured the upper bound in the comparison. Since missing frames were quite common, the effect of this can have been quite strong. We have illustrated this in the bottom row of Table 3, by showing what would have been the upper bound in the other extremity, where all missing OptiTrack frames had instead been interpreted as detecting no people.

3.5.2 Performance of face-detection based localization

We use the Viola Jones Face Detector module to detect faces from the input video stream. The detected face is then tracked by the face tracking module, which is based on the Direct Incremental Kernel-PCA Tracker (DIKT) proposed in [1]. The advantage of the algorithm is that the tracker is robust against illumination change, partial occlusion (including shadows casted onto the face) and large pose variation.

Based on the face tracking result, we then use a facial point tracker to track the 2D and 3D location of the 66 facial landmarks depicted in Figure 6. In addition, the facial point tracker also detects the face's 3D pose in terms of pitch, yaw and roll with respect to the camera's optical axis. Note that the face detection result is also sent to the facial point tracker to serve as a fall-back option in case the face tracker falls. The facial point tracker implements the Constrained Local Model (CLM) [2] - based Discriminative Response Map Fitting (DRMF) method proposed in [3].

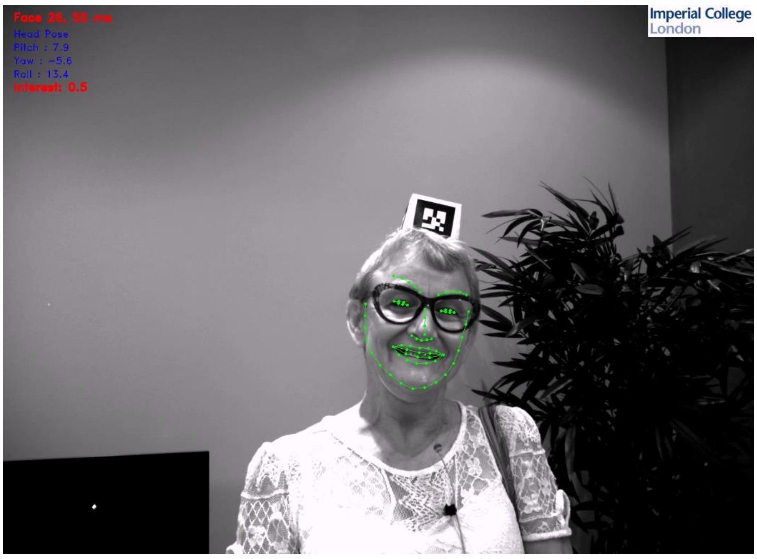

Figure 6 - Example of tracking a face from the data collected

Limitations: At the moment this module only tracks the most-dominant face in the scene. In addition, the localisation of the subjects is done with respect to the image coordinates frame and at this stage it cannot be combined with any other localisation components. However, these issues will be resolved in the final version as explained below.

4 Conclusions

In this document, we have discussed the placement of the Kinect depth sensor and the Dalsa camera that are used for localization. We have introduced a large data set of the Teresa robot interacting with several groups of people. This data set contains well over six hours of Kinect data, with ground truth on the positions of everyone involved in the interaction.

Based on this data set, we have been able to show that in interactions with groups of people, the proposed horizontal mounting of the Kinect sensor ensures that more people are within sensor range than there would be with a vertical mounting.

In addition, we have also been able to show that (partial) occlusions are not that common (3.2% of the frames) in interactions between the Teresa robot and a group. One reason for this could be that people actively adapt their position to the robot, probably to improve their own view and visibility. These findings indicate that occlusions should not be the primary focus when trying to improve the performance of localization algorithms in the context of the Teresa project, in contrast to what we suggested in our aims.

We have applied the Kinect skeletal tracker to the collected data and found that, after it has had a frame with a full (non-occluded) frontal view of a person, it is quite effective in tracking that person. The Kinect skeletal tracker can detect at least one person in at least 68.7% of the cases in which at least one person is in sensor range. Though this is already a reasonable lower bound, as we have argued this percentage can in reality be much higher, since the frames missing from the data will probably have had a negative effect.

To further improve on the localization provided by this system, for the next iteration of this deliverable, we aim to investigate ways in which the performance of the localization could be improved further. One approach to do so is to expand the used features from localization to also include those that could be more directly relevant such as orientation of upper body/head. An interesting addition would be derived features, such as who is part of a group (e.g. based on body posture info). The collected data set will help us train and test these algorithms.

We have also discussed a system for localization based on face detection. This system can currently only track single people and will for the next iteration of this deliverable be extended in order to be able to track multiple people. This will be done by running multiple face detectors in parallel and tracking the corresponding faces with multiple facial point trackers. In addition, a face verification system will be developed since it is essential for the robot to know if it interacts with the same person or a new person has joined the interaction. This is important since the camera used has a limited field of view, therefore it is not uncommon for a person to move outside of the camera’s view range and move back inside again during the same dialogue session.

In addition, we will aim to calibrate the camera so that we may map the 2D face locations obtained from the tracker to angular directions with respect to the robot coordinate system. This will be done by using the camera together with the stationary camera mounting on the robot. This will allow us to combine the information derived from the Kinect sensor with that of the face detection, to improve the performance and reliability of the localization module.

Overall, we have introduced suitable algorithms for localization and proposed various ways in which we could and aim to improve on those.

5 References

[1] S. Liwicki, S. Zafeiriou, G. Tzimiropoulos and M. Pantic, “Efficient Online Subspace Learning with an Indefinite Kernel for Visual Tracking and Recognition”, In: IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 10, pp. 1624-1636, October 2012.

[2] Saragih, Jason M., Simon Lucey, and Jeffrey F. Cohn. “Deformable Model Fitting by Regularized Landmark Mean-Shift”, International Journal of Computer Vision 91, no. 2 (2011): 200-215.

[3] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, “Robust Discriminative Response Map Fitting with Constrained Local Models”, 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013). Portland, Oregon, USA, June 2013.