Deliverable D7.2 Data Specification

Project acronym: *TERESA*

Project full title: Telepresence Reinforcement-Learning Social Agent

Grant agreement no.: 611153

Due-Date:	M6
Delivery:	M6
Lead Partner:	Imperial College London
Dissemination Level:	PU
Status:	Submitted
Version:	Final

DOCUMENT INFO

Date and version number	Author	Comments
20.05.2014 v01	Stavros Petridis	Initial Version
22.05.2014 v02	Kyriakos Shiarlis/Noe Perez Higueras/Jered Vroon/Khiet Truong	Adding sections
26.05.2014 v03	Jie Shen	Content Check
27.05.2014 v04	Stavros Petridis	Final Version

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

1 Executive summary

This deliverable D7.2 Data Specification elaborates on the data sets that will be collected during the TERESA project and the software that will be developed. It aims to provide information about the data collection procedure and the software modules in an easy-to-understand manner. A detailed plan about the experiments that will be conducted and the data that will be collected is presented. Details on how such data sets will be made available to the scientific community within ethics guidelines and regulations are also provided. Finally, a description about the software that will be developed based on the collected data set is provided. Naturally, This deliverable is organised in 3 sections: 1) description of the experiments and data collection, 2) software modules to be developed, 3) organisation and dissemination of the data set to the scientific community.

Dissemination of open source software and data sets will be coordinated by Imperial College London and UvA, the project coordinator, working together with the other partners who will develop software modules and MADoPA who will provide the facilities for conducting the experiments and collecting the required data. Estimated indicative person-months is 6.

2 Contributors

The authors of the deliverable are Noe Perez Higueras, Jered Vroon, Khiet Truong, Kyriakos Shiarlis, Jie Shen, Stavros Petridis, and Maja Pantic. Jered Vroon contributed the description of the interaction episodes (Section 3.1.1) and together with Khiet Truong provided the description of the software that will be developed by the University of Twente (Section 3.2.2). Noe Perez Higueras and Kyriakos Shiarlis provided the description of the software that will be developed by UPO (Section 3.2.3) and the University of Amsterdam (section 3.2.4), respectively. Jie Shen and Stavros Petridis coordinated the deliverable and wrote the first version. Maja Pantic provided feedback on the deliverable. On-going work in relation to the activities described in this deliverable will be reported in D7.3 (Exploitation Plan).

3 Data sets and software specification

3.1 Experiments and Data Collection

The TERESA project aims to develop the modules required for a telepresence robot to display socially appropriate behavior in various interactions with groups of or individual elderly. Because it would be unfeasible to investigate all possible such interactions, this document aims to specify a challenging, relevant and manageable subset. To do so, we have grouped these interactions in practical and meaningful episodes. They were based on the description of work in the grant agreement and the requirements report (D6.1).

3.1.1 Interaction Episodes for TERESA

The episodes are defined on an abstract level as they are intended as the basis for our scenarios and experiments throughout the project. The scenarios will consist of real-world implementations of these episodes in a meaningful context (e.g. the philosophy café) such that the interactions described in the episode are both necessary and sufficient.

We here describe three episodes; social navigation, social conversing and free-roaming. Furthermore, we describe their variables (e.g. is the interaction target an individual or a group?) and the different combinations of values for these variables. To prevent combinatorial explosion, we have also included suggestions for particular such combinations to be excluded from our investigations. An overview of the terms used in the description of the episodes below is given in Table I.

In addition, in our first experiments we will consider three types of visitors:
Visitor 1: Physically present Person: In this case, the visitor will physically walk through the facility and execute an episode, with no TERESA Robots or Controllers involved. This will serve as a baseline comparing between telepresence and actual presence. Subtle behaviors of the TERESA Robot (e.g. precise orientation, head tilt for social gestures) may be hard for the pilot to perform, though they could be easy for the semi-autonomous version. The data collected with the physically present person may help recognize these behaviors. This type of visitor will, unlike the other types, not come equipped with a wide range of sensors. For this kind of visitor, environmental sensors (e.g. using visual markers) will thus be essential.
Visitor 2: Pilot (tele)present through a TERESA Robot: In this case, a pilot will control a TERESA Robot as it moves through the facility and executes an episode.
Visitor 3: Pilot (tele)present through a TERESA Robot with a confederate performing the actual navigation: This type of Visitor is similar to the previous one, with the exception that instead of the Pilot, a confederate performs all navigation actions. This type of visitor will serve two important purposes. First, we would expect that because the confederate can dedicate all attention to the navigation, she can display more complicated social interactions. Second, this set-up could effectively surmount to a Wizard-of-Oz study of the semi-autonomous TERESA Robots this project aims to develop. As such, we can use it to evaluate how the Pilot experiences transferring some autonomy to the system.

Term	Explanation
TERESA Robot	A Giraff supplemented with extra computers and sensors
TERESA Controller	A computer running Giraff software and supplemented with a camera and headset, which can be used to control the TERESA Robot
Pilot	A human being who sits at a TERESA Controller and controls a TERESA robot. We will assume in each episode that the pilot has had some basic practice with using the control interface and can do so with some skill
Interaction target	People in the same room with the TERESA Robot who interact with it
Visitor	Either a pilot that uses the TERESA robot to visit the test facility and interact with the interaction target, or a physically present human that does the same
Confederate	Actor that is involved in running the experiment and not a normal subject
Interaction	Interaction between a visitor and an interaction target. In the context of the Teresa project we will be primarily interested in interactions that require some sort of socially appropriate behavior from the visitor
Episode	A sequence of interactions wherein a given visitor interacts with an interaction target
Scenario	Real-world implementation of an episode in a meaningful context such that the interactions described in the episode are both necessary and sufficient
Experiment	Controlled execution of an episode in the context of a scenario under different conditions, aimed at acquiring data

Table 1: Explanation of terms used in this deliverable.

3.1.1.1 Episode 1 : Social Navigation

For this episode, we have aimed at creating specific conflict situations for navigation where some sort of social behavior would be required. To ensure that these conflict situations arise repeatedly, we propose to use confederates as interaction targets. As a result, this episode will probably be the most controlled of the three described in this document.

Set-up

In this episode, the visitor will navigate from A to B, encountering a variable obstacle on the way. These obstacles will be interaction targets standing/moving such that the navigation of the visitor will probably have to take them into account properly to be perceived as social.

Variable values

The variables in this set-up are the kind of interaction target and their direction. We propose the following values for these variables;

Interaction target:	[individual / group (2-4 people)]
Interaction target direction:	[coming to you / stationary]

What behavior is social in these situations, will probably also depend on the possibility to navigate around the interaction target. We propose to look at two specific situations where such circumnavigation is impossible. In the first the interaction target is an individual coming to the visitor such that the conflict occurs in a narrow passage (e.g. a doorway). In the second, the interaction target is a stationary group standing in a narrow passage (e.g. a narrow hallway) such that the visitor must pass through the group.

Proposed combinations of variables values

1. Passing by an individual coming to the visitor, circumnavigation possible (e.g. in a hallway)
2. Passing by an individual coming to the visitor, circumnavigation impossible (e.g. in a doorway)
3. Passing by a stationary individual, circumnavigation possible
4. Passing by a group coming to the visitor, circumnavigation possible

5. Passing by a stationary group, circumnavigation possible (i.e. visitor can pass around group)
6. Passing by a stationary group, circumnavigation impossible (i.e. visitor must pass through group)

3.1.1.2 Episode 2 : Social Conversation

Social conversation is aimed at the whole range of behavior required for a social conversation – including the approach to and retreat from the interaction target.

Set-up

In this episode the visitor will go through three phases;

1. *Approach*: Approach interaction target
2. *Converse*: Converse with the interaction target
3. *Retreat*: Retreat from the interaction target

The conversation in this episode can/should be such that it elicits both positive and negative responses from those present. This will ensure a more diverse data set. To give us a handle with which we can control this, the visitor and interaction target will be instructed on the topic for their conversation.

Variable values

Interaction target:	[individual / group (2-4 people)]
Interaction target direction:	[sitting / standing / moving while conversing]

Conversation topic:	[neutral / loaded]
Conversation instruction:	[congruent / incongruent]

Here ‘moving while conversing’ will focus on the situation where the interaction target is standing and engaged in an activity involving substantial changes in orientation and pose as well as some moving around within a room (e.g. to direct the visitor to something). Furthermore, because we consider it a less likely and rather complicated setting, we propose not to investigate moving while conversing with groups.

We propose two approaches to ensuring that the conversation in this episode elicits both positive and negative responses. The first approach is picking a loaded conversation topic. Though the topic should be such that it could elicit an emotional response, for moral and ethical reasons it should not be confronting or unpleasant. Possible loaded topics could be quality of the food at the nursing home or sport teams. Which topics are suitable will be dependent on the scenario and perhaps even on the individual subjects. The second approach involves giving the visitor and the interaction target incongruent instructions on the conversation topic. For example, we could tell the pilot to talk about the weather and the interaction target to talk about food. We expect the incongruence will elicit some sort of emotional response.

We chose these two approaches because they will probably influence the kind of observed responses, but still allow the conversation to be rather natural. From our preliminary results we should decide if these approaches effectively produce the desired effect. Because ‘over-use’ of the incongruent instructions will probably influence its effect, we propose not to look at the combination of loaded conversation topics and incongruent instructions.

Proposed combinations of variables values

1. Social conversation with an individual who is sitting
2. Social conversation with an individual who is standing
3. Social conversation with an individual who is moving while conversing
4. Social conversation with a group that is sitting
5. Social conversation with a group that is standing

The combinations of conversation topics and instructions that will be investigated are the following:

1. Neutral conversation topic, congruent instructions
2. Neutral conversation topics, incongruent instructions
3. Loaded conversation topic, congruent instructions

3.1.1.3 Episode 3 : Free-roaming

This last episode is aimed at allowing the visitor to behave as normally as possible. So rather than requiring any specific behavior, in this episode the visitor is simply allowed to interact freely with one or more of various possible interaction targets. This episode can focus on the

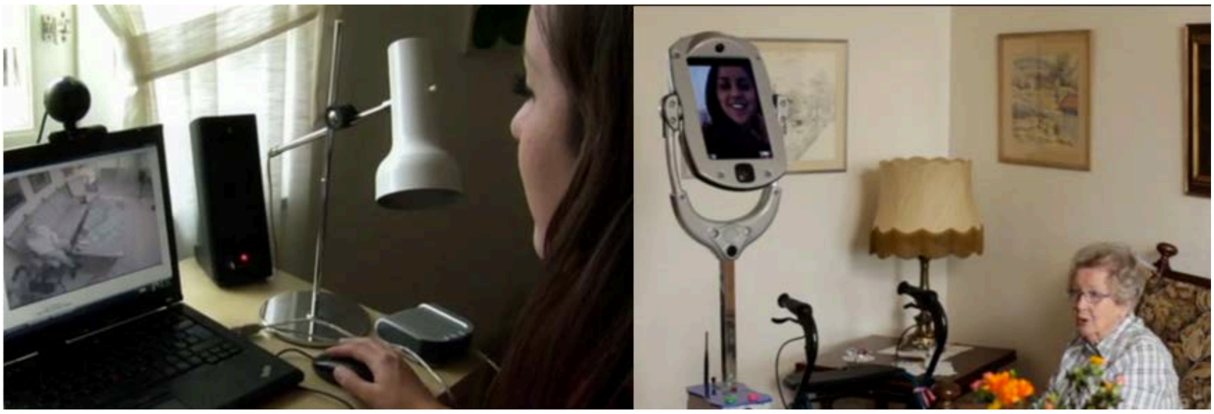


Figure 1: Example of Telepresence

cases where the visitor is a (tele)present Pilot, excluding the case where the visitor is a physically present human. An example scenario that could implement this episode is that in which a visiting pilot is asked to use the Teresa robot to interact with attendees of a Philosophy café meeting in any way.

We believe it is important to include this episode because it might allow us to identify other interactions not included in this document but essential for a proper autonomous social functioning of the robot. Furthermore, we expect the more open assignment may result in more natural behavior, allowing us to control that the instructions in the other episodes do not have a disturbing effect on the kinds of behaviors displayed.

3.1.2 Data Collection

As mentioned in the previous section in all the experiments there will be a pilot who will control the TERESA robot through the TERESA controller as shown in Figure 1. During the execution of each experiment the reaction of the pilot and the individual or group of individuals interacting with the robot will be recorded using the equipment mounted on the TERESA controller and the TERESA robot, respectively. More specifically, a high-resolution camera on the TERESA controller will record the pilot's facial expressions and upper-body movements. In addition, a microphone on the controllers's side will record the pilot's vocalisations. The TERESA robot will also be equipped with a camera, as shown in Figure 2, and a microphone in order to record the facial expressions, upper-body movements and vocalisations of the people in the robot's vicinity. Laser scanners mounted on the robot will be used to locate the robot in the scenario, to avoid obstacles (either static or dynamic) and to locate and detect people around the robot. The lasers rangefinders used in TERESA can have a range up to 10 meters. A Kinect will also be used to localise and estimate the body pose of people within a range of 1.5m and 4.5m from the robot.

During the execution of each experiment all these data streams will be recorded and will be made available online, as explained in section 3.4, within ethics guidelines and regulations. All participants will be asked to give their consent for their data to be used for reserach purposes and possibly for scientific publications/presentations.

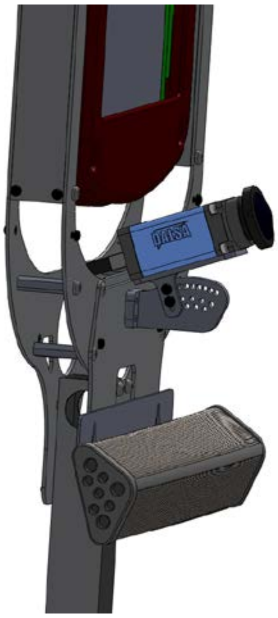


Figure 2: TERESA robot with mounted camera.

3.2 Software specification

Several software modules will be developed by the partners for each work package and their integration with the TERESA robot will be done by our industrial partner Giraff. All the software modules will be available for the scientific community via the project's website. Details about the development, the programming languages that will be used and the software distribution can be found in the following sections.

3.2.1 Software developed by Imperial College London for WP2

Imperial College London will develop software that aims to understand the reactions of the pilot and of people in the robot's vicinity. More specifically the software will:

1) Detect multiple faces in a scene with at least 80% accuracy and track those faces over time regardless of clutter and severe lighting conditions.
2) Recognise facial expressions and nonverbal cues like smiles, frowns and head nods/shakes.
3) Recognise positive/negative valance and low arousal.

The system structure is illustrated in Figure 3. As shown in the figure, the system consists of multiple interconnected modules, which will be integrated into a single piece of software using the HCI^2 Framework [1].

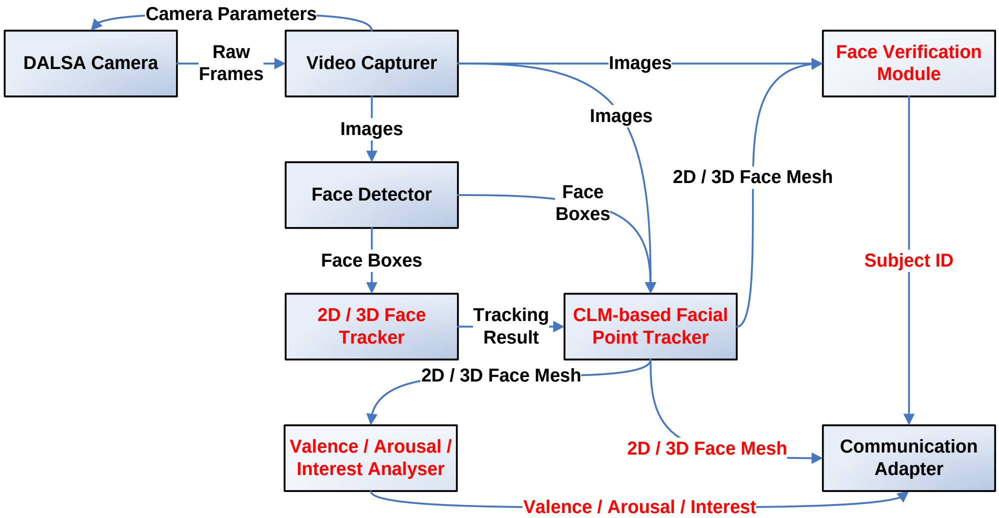


Figure 3: Overview of the software modules that will be developed for the TERESA project.

We will use the Viola Jones Face Detector module to detect faces from the input video stream. The detected faces will then be tracked by the face tracking module ('2D / 3D Face Tracker' in Figure 3), which is based on the Direct Incremental Kernel-PCA Tracker (DIKT) proposed in [2]. 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 will then use a facial point tracker ('CLM-based Facial Point Tracker' in Figure 3) to track the 2D and 3D location of 66 facial landmark points. 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. Figure 4 shows some combined result of face tracking and facial point tracking on the dataset recorded in the FROG project. The existing framework will be extended to be able to cope with detecting and tracking multiple faces under clutter and severe lighting conditions.

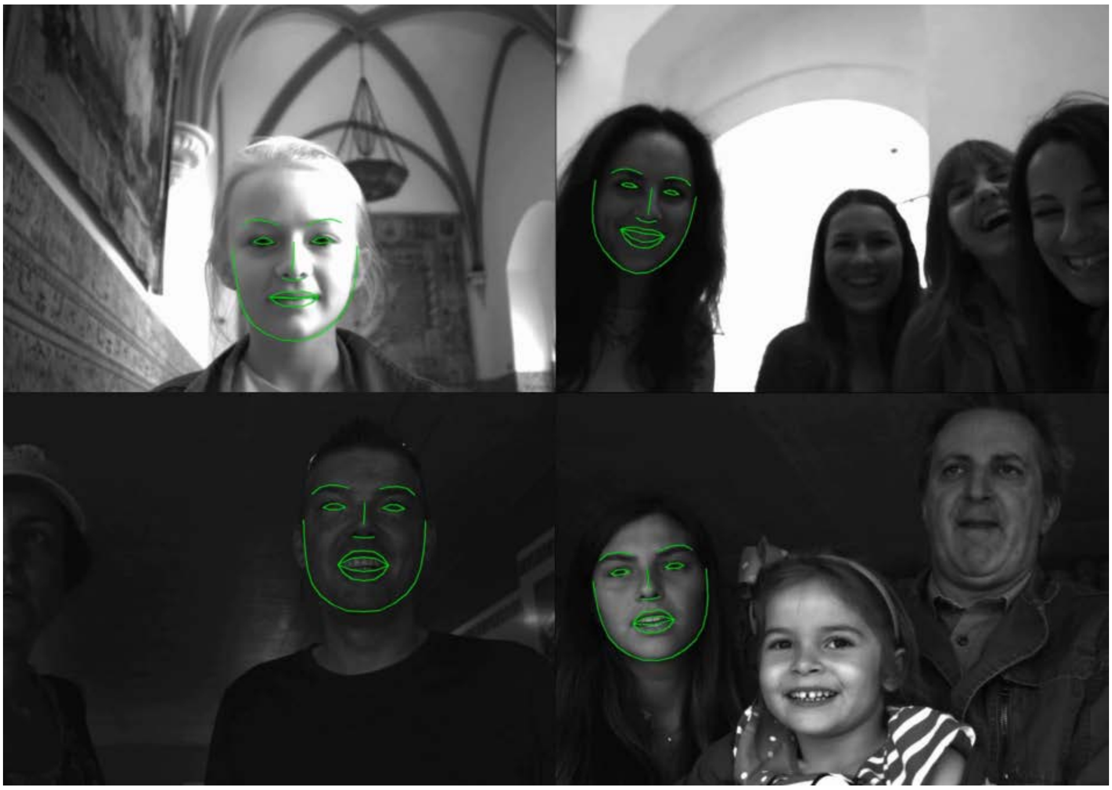


Figure 4: Facial point (landmark) tracking results on the dataset recorded in the FROG project. The existing framework will be extended to be able to cope with detecting and tracking multiple faces under clutter and severe lighting conditions.

The valance / arousal / interest analyser will predict the person's emotional state using the face mesh provided by the facial point tracker. By applying the algorithms proposed in [3] and [4], the analyser will output continuous prediction results of the person's emotional state in terms of valance, arousal and level of engagement. The actual analyser will be trained on existing datasets and on the dataset recorded within the TERESA project**.** The face tracker and the facial point tracker will be used to track the facial landmarks of all the faces detected in the video clips. The tracking results will be used as the feature vector for the analyser. Since the analyser works with low-dimensional sparse features, the computational cost of the prediction step is relatively. Therefore, instead of developing the module entirely in C/C++, we will implement the module as a Matlab programme wrapped within a C/C++ shell using the Matlab Engine API [5] and the HCI^2 Framework SDK to handle data communication between the core algorithm implementation and other modules within the HCI^2 Framework.

Finally a face verification module will be built to give a unique ID to each person being analysed. Since the camera used by the TERESA robot has a limited field of view, it will not be uncommon for a person to move outside of the camera's view range and move back inside again during the same dialogue session. In this case, it is essential for the robot to know that it is still interacting with the same person. Since the face tracker and the facial pointer tracker cannot keep tracking the face when the person is temporarily outside of the camera's view, we cannot assume every new tracking session corresponds to a new person. Instead, we will use the face verification module to check whether the newly tracked face belongs to a previously unseen person, and if not, give the tracking result the person's old ID. Instead of performing face recognition, we will derive a verification scheme based on the face alignment algorithm proposed in [6]. Specifically, the face verification module will capture a picture of the person's face when he / she is detected by the robot for the first time. When a new face tracking session starts, we will try to align the new face to every existing face image stored in the module's memory. The alignment error will be used to determine which person the tracked face belongs to, or if the face belongs to an un-seen person (when the error exceeds a pre-defined threshold).

3.2.2 Software developed by University of Twente for WP2

The University of Twente will develop software which aims to:

1) detect with some accuracy relevant social cues from the data provided by the robot-mounted Kinect sensor.
2) monitor continuous dialogue.

With respect to the first target, we will initially focus on the position of people within range, their orientation and their ordering into groups and roles within those groups. As our knowledge progresses, we will expand the system to derive further relevant social cues based on body postures. Inputs will thus primarily come from the robot-mounted Kinect sensor, though we may also use (1) information about the people detected with the laser range finder by the UPO and (2) the results of the head orientation and facial expression recognition from ICL. Outputs will be a range of socially relevant information about the detected people with an associated certainty. The algorithms will probably be written in C++, making use of the OpenCV libraries and either OpenKinect or the Kinect libraries provided by Microsoft.

With respect to the second target, we aim to develop algorithms that will interpret the raw acoustic features extracted from the speech signal with regards to the `social quality' of the conversation. Inputs will be acoustic features such as F0, energy, and turn-taking features.

Outputs will be an estimation of aspects of `quality' of the conversation: how much aligned/engaged are the interaction partners towards each other, how good is the conversation going, are the interaction partners `in sync'? We will use open source publicly available acoustic feature extraction programs such as Praat or openSMILE. The language of the software component will probably be Matlab or C++.

3.2.3 Software developed by University Pablo de Olavide for WP4

UPO will develop software that aims to:

1) help the robot navigate autonomously by avoiding obstacles.
2) allow the robot navigate taking into account socially normative behaviors.

The navigation stack is in charge of the navigation autonomy of the TERESA platform. From the high-level commands of the operator (like indicating the person he wants to talk to), the main objective of the module is to command motions to the robot to achieve the task, while avoiding obstacles and attaining a socially compliant motion behavior. The module will be developed in an iterative fashion following the different data recording sessions of the project.

The inputs for the navigation system will be:

- The front and rear lasers, for localization and obstacle detection (either static or dynamic).
- The gyroscopes from the IMU, for localization.
- People poses provided by UT from the Kinect data.
- Motors states in order to calculate the odometry.

The outputs will be:

- velocity commands for robot movement.
- Localization of the robot.

The social navigation stack, which is divided in two modules, global path planning and behavior control (local planner), will be programmed using C++ language and based on ROS (Robot Operating System).

The path planning module will obtain efficient and human-aware global paths. Human motion, spatial human activities, and human interaction models will be considered when computing the paths, as well as the learned cost functions from previous experiences. The planner should take into account also joint human-robot tasks, such as moving together with a partner while engaged into a conversation, as well as uncertainties (due to future human trajectories, inaccurate models, and non-observability of features) and social costs.

The behavior control module will be in charge of executing the planned path by the previous module, while adapting to the current situation using sensory feedback. This module will receive, thus, the computed path and all the sensorial information and will determine the final velocity commands to the robot. A local efficient path planner will be developed, based on learning, to achieve this task, while taking into account the socially normative behaviors. All the software employed by UPO will be open source.

3.2.4 Software developed by University of Amsterdam for WP5

The UvA will develop software that aims to:

1) Learn socially appropriate behaviour from the sensed environment. This consists of learning offline two cost functions, one for navigation and one for body pose.
2) Execute the learned behaviour online. The UvA will collaborate with UPO in order to use the cost functions to elicit the desired behaviour.

Off-line Learning: These modules will be developed in Python since fast prototyping is important and computational efficiency is less of an issue due to their off-line nature. Data will come from the experiments in MADOPA as well as features from the work of the other partners. Each cost function will consist of four modules that will use this data for learning:

1. A module that maps the pilot-related inputs to a cost and a confidence. We will develop semi-supervised learning algorithms as subroutines for its construction. Implementations of any new methods devised here will be made available in open source along with the data.

2. A module that maps the audience (people around the robot) inputs, to a cost and a confidence, in the same manner as 1.

3. A module that is learned from human demonstration using novel Apprenticeship Learning methods such as Inverse Reinforcement Learning (IRL) or Supervised Learning. As these methods have not seen much exposure to real-world scenarios yet, it is expected that new algorithms will be developed, whose implementation will be finally made public.

4. A global cost function that will assess the “general situation”. This will use the same methods as 1 and 2 but will incorporate labels received from an annotator having access to the whole dataset (pilot and audience side). Again, it will output a cost and confidence.
5. An integrator that will be able to intelligently weigh these costs using principles of Multi Objective Planning. This will output the final cost for the set of input features. Implementations of any advancements to the state of the art in cost function integration will be made public.

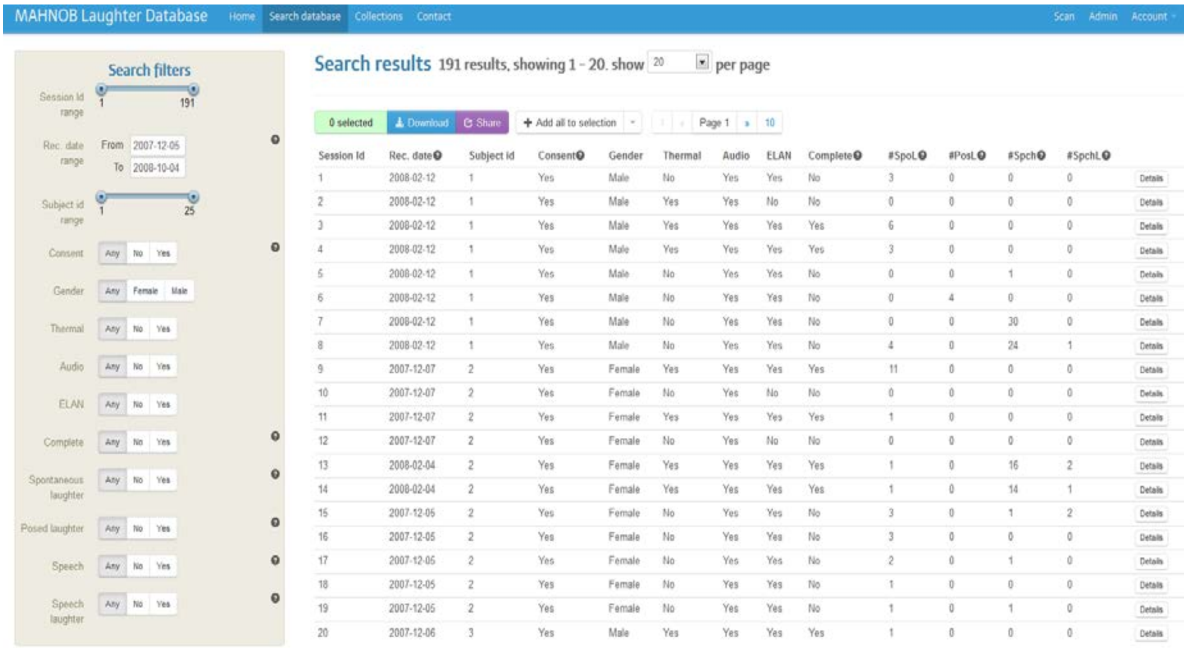
In addition to the cost functions, we will develop an active perception module. This will be responsible for trading of some social behaviour for the sake of better sensing of the environment.

On-Line modules: Firstly, the cost functions developed offline will be coded onto the onboard computer. We will also build modules that make decisions based on these cost functions. More specifically, the UvA will build software that will plan and execute series of appropriate body poses based on the sensed state of the world. This will be done in C++ using ROS premises. It is also expected that we will collaborate with UPO (WP4) in order to develop the software by which navigation, active perception and body pose communicate. These modules will take as input the sensed state of the environment and make high level decisions on whether to navigate, or change body pose

3.3 Database

The collected data set will be freely available online to the scientific community for research purposes. The database will be organized in sessions, possibly corresponding to the execution of one episode, and within each session there will be several tracks containing the sensor data. There will be at least five different tracks per session: two video and two audio streams (one for the pilot and one for the robot), and Kinect data recorded by the TERESA robot. If necessary additional data recorded by the TERESA robot can be included, like laser scanners data.

To facilitate easy browsing of the database, low-quality previews of all tracks will be available on the web interface. The web-based interface of the database will have a search form functionality that will allow the user to search for the exact set of data needed. Search options will include, for example, duration of the recordings, whether a data track type is available, whether subjects gave written consent for publications and presentations, interaction type all the attributes of each data track, e.g. frame rate and size for video, sampling rate for audio etc. An example of how the web-interface of the database will look like is shown in Figure 4.



Search filters

Session Id range: 1 – 191

Rec. date range: From 2007-12-05 To 2008-10-04

Subject id range: 1 – 25

Consent: Any No Yes

Gender: Any Female Male

Thermal: Any No Yes

Audio: Any No Yes

ELAN: Any No Yes

Complete: Any No Yes

Spontaneous laughter: Any No Yes

Posed laughter: Any No Yes

Speech: Any No Yes

Speech laughter: Any No Yes

Search results 191 results, showing 1 - 20. show 20 per page

Session Id	Rec. date	Subject id	Consent	Gender	Thermal	Audio	ELAN	Complete	#SpoL	#PosL	#Spch	#SpchL	
1	2008-02-12	1	Yes	Male	No	Yes	Yes	No	3	0	0	0	Details
2	2008-02-12	1	Yes	Male	Yes	Yes	No	No	0	0	0	0	Details
3	2008-02-12	1	Yes	Male	Yes	Yes	Yes	Yes	6	0	0	0	Details
4	2008-02-12	1	Yes	Male	Yes	Yes	Yes	Yes	3	0	0	0	Details
5	2008-02-12	1	Yes	Male	No	Yes	Yes	No	0	0	1	0	Details
6	2008-02-12	1	Yes	Male	No	Yes	Yes	No	0	4	0	0	Details
7	2008-02-12	1	Yes	Male	No	Yes	Yes	No	0	0	30	0	Details
8	2008-02-12	1	Yes	Male	No	Yes	Yes	No	4	0	24	1	Details
9	2007-12-07	2	Yes	Female	Yes	Yes	Yes	Yes	11	0	0	0	Details
10	2007-12-07	2	Yes	Female	No	Yes	No	No	0	0	0	0	Details
11	2007-12-07	2	Yes	Female	Yes	Yes	Yes	Yes	1	0	0	0	Details
12	2007-12-07	2	Yes	Female	No	Yes	No	No	0	0	0	0	Details
13	2008-02-04	2	Yes	Female	Yes	Yes	Yes	Yes	1	0	16	2	Details
14	2008-02-04	2	Yes	Female	Yes	Yes	Yes	Yes	1	0	14	1	Details
15	2007-12-05	2	Yes	Female	No	Yes	Yes	No	3	0	1	2	Details
16	2007-12-05	2	Yes	Female	No	Yes	Yes	No	3	0	0	0	Details
17	2007-12-05	2	Yes	Female	No	Yes	Yes	No	2	0	1	0	Details
18	2007-12-05	2	Yes	Female	No	Yes	Yes	No	1	0	0	0	Details
19	2007-12-05	2	Yes	Female	No	Yes	Yes	No	1	0	1	0	Details
20	2007-12-06	3	Yes	Male	Yes	Yes	Yes	Yes	1	0	0	0	Details

Figure 5: The MAHNOB Laughter Database. The data set recorded in the TERESA project will be organised in the same way and will have a similar web interface which will allow the search of the database based on specified attributes.

4 Conclusions

Data collection will be a very important task in the TERESA project since most of the algorithms and software that will be developed will be based on these data sets. The goal is to record data following the specifications described in this deliverable and make them available to the scientific community through a web-based searchable interface. In addition, open source software that will be developed during the project will also be made available to the scientific community through the project's website.

5 Annex

This section describes the specifications of the data streams that will be stored to the database.

Camera: A Dalsa Genie HM1400 / XDR camera is used for video capturing. Through GigE interface, the camera is capable of capturing grayscale video frames at 64.00 FPS, with a spatial resolution of 1400x1024 pixels.

Microphone: what will be stored in the database are mono 16kHz pcm wav files.

Kinect: Color and depth information from the Kinect will be stored using the default Kinect formats (RGB + depth (up to 8m), 640x480, 30fps) – though we may add some compression (e.g. dropping half of the frames) to keep the amount of data within check. Additionally, the audio stream from the Kinect sensor may be used in a later stadium as well.

Laser Scanners: Raw data at frequency of 40 Hz.

Other potential data streams: IMU raw data (~100 Hz), Velocity commands sent to the robot (~20 Hz), Odometry data calculated by UPO driver (15~20Hz).

6 References

[1] J. Shen and M. Pantic, “HCI^2 Framework: A Software Framework for Multimodal Human-Computer Interaction Systems”. In: IEEE Transactions on Cybernetics, Vol. 43, Issue 6, pp. 1593-1606, December 2013.

[2] 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.

[3] M. A. Nicolaou, H. Gunes, M. Pantic, “Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space”, IEEE Transactions on Affective Computing. pp. 92 - 105, 2011

[4] M. A. Nicolaou, H. Gunes, M. Pantic, “Output-associative RVM regression for dimensional and continuous emotion prediction”, Image and Vision Computing, Special Issue on The Best of Automatic Face and Gesture Recognition 2011. 30: pp. 186 - 196, Issue 3. 2012.

[5] “Introducing MATLAB Engine”. [Online]. Available: http://www.mathworks.co.uk/help/matlab/matlab_external/introducing-matlab-engine.html.

[6] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, “Robust and Efficient Parametric Face Alignment”, Proceedings of IEEE Int’l Conf. on Computer Vision (ICCV 2011). pp. 1847 - 1854. November 2011.

7 Abstracts

[1] J. Shen and M. Pantic, “HCI^2 Framework: A Software Framework for Multimodal Human-Computer Interaction Systems”. In: IEEE Transactions on Cybernetics, Vol. 43, Issue 6, pp. 1593-1606, December 2013.

Abstract - This paper presents a novel software framework for the development and research in the area of multimodal human-computer interaction (MHCI) systems. The proposed software framework, which is called the HCI^2 Frame work, is built upon publish / subscribe (P/S) architecture. It implements a shared-memory-based data transport protocol for message delivery and a TCP-based system management protocol. The latter ensures that the integrity of system structure is maintained at runtime. With the inclusion of ‘bridging modules’, the HCI^2 Framework is interoperable with other software frameworks including Psyclone and ActiveMQ. In addition to the core communication middleware, we also present the integrated development environment (IDE) of the HCI^2 Framework. It provides a complete graphical environment to support every step in a typical MHCI system development process, including module development, debugging, packaging, and management, as well as the whole system management and testing. The quantitative evaluation indicates that our framework outperforms other similar tools in terms of average message latency and maximum data throughput under a typical single PC scenario. To demonstrate HCI^2 Framework’s capabilities in integrating heterogeneous modules, we present several example modules working with a variety of hardware and software. We also present an example of a full system developed using the proposed HCI^2 Framework, which is called the CameGame system and represents a computer game based on hand-held marker(s) and low-cost camera(s).

[2] 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.

Abstract - We propose an exact framework for online learning with a family of indefinite (not positive) kernels. As we study the case of nonpositive kernels, we first show how to extend kernel principal component analysis (KPCA) from a reproducing kernel Hilbert space to Krein space. We then formulate an incremental KPCA in Krein space that does not require the calculation of preimages and therefore is both efficient and exact. Our approach has been motivated by the application of visual tracking for which we wish to employ a robust gradient-based kernel. We use the proposed nonlinear appearance model learned online via KPCA in Krein space for visual tracking in many popular and difficult tracking scenarios. We also show applications of our kernel framework for the problem of face recognition.

[3] M. A. Nicolaou, H. Gunes, M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space", IEEE Transactions on Affective Computing. pp. 92 - 105, 2011

Abstract - Past research in analysis of human affect has focused on recognition of prototypic expressions of six basic emotions based on posed data acquired in laboratory settings. Recently, there has been a shift towards subtle, continuous, and context-specific interpretations of affective displays recorded in naturalistic and real-world settings, and towards multi-modal analysis and recognition of human affect. Converging with this shift, this paper presents, to the best of our knowledge, the first approach in the literature that: (i) fuses facial expression, shoulder gesture and audio cues for dimensional and continuous prediction of emotions in valence and arousal space, (ii) compares the performance of two state-of-the-art machine learning techniques applied to the target problem, the bidirectional Long Short-Term Memory neural networks (BLSTM-NNs) and Support Vector Machines for Regression (SVR), and (iii) proposes an output-associative fusion framework that incorporates correlations and covariances between the emotion dimensions. Evaluation of the proposed approach has been done using the spontaneous SAL data from 4 subjects and subject-dependent leave-one-sequence-out cross-validation. The experimental results obtained show that: (i) on average BLSTM-NN outperform SVR due to their ability to learn past and future context, (ii) the proposed output-associative fusion framework outperforms feature-level and model-level fusion by modeling and learning correlations and patterns between the valence and arousal dimensions, and (iii) the proposed system is well able to reproduce the valence and arousal ground truth obtained from human coders

[4] M. A. Nicolaou, H. Gunes, M. Pantic, “Output-associative RVM regression for dimensional and continuous emotion prediction”, Image and Vision Computing, Special Issue on The Best of Automatic Face and Gesture Recognition 2011. 30: pp. 186 - 196, Issue 3. 2012.

Abstract - Many problems in machine learning and computer vision consist of predicting multi-dimensional output vectors given a specific set of input features. In many of these problems, there exist inherent

temporal and spatial dependencies between the output vectors, as well as repeating output patterns and input – output associations, that can provide more robust and accurate predictors when modeled properly. With this intrinsic motivation, we propose a novel Output-Associative Relevance Vector Machine (OA-RVM) regression framework that augments the traditional RVM regression by being able to learn non-linear input and output dependencies. Instead of depending solely on the input patterns, OA-RVM models output covariances within a predefined temporal window, thus capturing past, current and future context. As a result, output patterns manifested in the training data are captured within a formal probabilistic framework, and subsequently used during inference. As a proof of concept, we target the highly challenging problem of dimensional and continuous prediction of emotions,

and evaluate the proposed framework by focusing on the case of multiple nonverbal cues, namely facial expressions, shoulder movements and audio cues. We demonstrate the advantages of the proposed OA-RVM regression by performing subject-independent evaluation using the SALdatabase that constitutes naturalistic conversational interactions. The experimental results show that OA-RVM regression outperforms the traditional RVM and SVM regression approaches in terms of accuracy of the prediction (evaluated using the Root Mean Squared Error) and structure of the prediction (evaluated using the correlation coefficient), generating more accurate and robust prediction models.

[6] G. Tzimiropoulos, S. Zafeiriou, M. Pantic, “Robust and Efficient Parametric Face Alignment”, Proceedings of IEEE Int’l Conf. on Computer Vision (ICCV 2011). pp. 1847 - 1854. November 2011.

Abstract - We propose a correlation-based approach to parametric object alignment particularly suitable for face analysis applications which require efficiency and robustness against occlusions and illumination changes. Our algorithm registers two images by iteratively maximizing their correlation coefficient using gradient ascent. We compute this correlation coefficient from complex gradients which capture the orientation of image structures rather than pixel intensities. The maximization of this gradient correlation coefficient results in an algorithm which is as computationally efficient as ℓ2 norm-based algorithms, can be extended within the inverse compositional framework (without the need for Hessian re-computation) and is robust to outliers. To the best of our knowledge, no other algorithm has been proposed so far having all three features. We show the robustness of our algorithm for the problem of face alignment in the presence of occlusions and non-uniform illumination changes. The code that reproduces the results of our paper can be found at http://ibug.doc.ic.ac.uk/resources.