CoCoMaps Project

CMLabs | IIIM

CoCoMaps Demo-3 Report

Demonstration date: 26 March 2018

Link to video on Youtube: https://youtu.be/Xb492moHr20

INTRODUCTION

The CoCoMaps Project is a joint effort by CMLabs (UK) and the Icelandic Institute for Intelligent Machines (Iceland) to integrate an architecture for robot control and interaction with humans using natural communication.

This report describes data from Demo-3 of the CoCoMaps project.

The goal of Demo-3 is to demonstrate the advances made throughout this project in the development of the Collaborative Cognitive Map Architecture, including communication with multiple humans, information extraction, and more. It is listed as Milestone 7 in the CoCoMaps including the following deliverables:

- **T12.D1** Four-way Turn-Taking
- **T13.D1** Demo 3: Collaborative Information Extraction

This report describes the successful conclusion of Demo-3. It includes data showing how the robots collaborate, negotiate skills, and interact with two humans, in concert with other functionalities of CoCoMaps (including the skills previously demonstrated in Demo-1 and Demo-2) integrated into a running system involving two robots with two humans present in the robots’ area of operation.

The results of processing times, CPU loads, and overall architecture reliability are shown to be within target ranges of the project as a whole.

Demo-3 Goals. Demo-3 aims at demonstrating multi-party collaboration capabilities using dialog, integrated with navigation and appropriate visual competencies, virtual control panel interaction, where two robots work in an environment extracting directions from humans. Specifically, the robots collaborate and communicate with each other and with two humans, to perform a task initiated by the humans. Demo-3 subsumes Demo-2 and differs from it in that the collaboration involves extracting the specific task, sub-tasks, and individual actions dynamically from two humans during the session via 4-way interaction. Task- and role assignment is done by the robots dynamically during the interaction. As in Demo-2 the communication and dialog acts are both time- and context-dependent.

We test the ability of the system to do this by running specific scenarios designed for that purpose. To ensure consistency and data reliability we run similar scenarios several times in the same area.

Demo-3 demonstrates the successful design and implementation of the CCM to facilitate collaboration between two robots and two humans using information extraction, interactive dialogue and dynamically negotiated task assignments and roles.

Table 1.
KPIs from CoCoMaps proposal relevant to Demo-3.

1	Ability of current state of the art running on one computer	M10		One computer able to see, listen and speak in simple setup	Video recording, statistics graphs
2	Ability of real-world robot-robot interaction using new collaborative CMArch	M13	One Turtlebot able to see, listen and speak in simple setup	Two Turtlebots able to communicate via CMArch	Video recording, statistics graphs
3	Ability of real-world multi-robot-human interaction using collaborative CMA and speech	M15	Two Turtlebots able to communicate via CMA	Two Turtlebots able to communicate with one human via CMA and voice	Video recording, statistics graphs
4	Efficiency of collaborative detection of humans	M16	Initial measurement of detection efficiency at current SOA implementation	Measurement of detection efficiency at Demo 1	Measure added efficiency (speed, effort, error rate) of collaborative detection
5	Efficiency of collaborative tracking of humans	M16	Initial measurement of tracking efficiency at current SOA implementation	Measurement of tracking efficiency at Demo 1	Measure added efficiency (speed, effort, error rate) of collaborative tracking
6	Efficiency of collaborative informatio n extraction through dialogue	M17	Initial measurement of extraction efficiency at current SOA implementation	Measurement of extraction efficiency at Demo 3	Measure added efficiency (speed, effort, error rate) of collaborative extraction
7	Efficiency of collaborative task extraction through dialogue	M18	Measurement of extraction efficiency at Demo 2	Measurement of extraction efficiency at Demo 3	Measure added efficiency (speed, effort, error rate) of collaborative extraction
8	Real-time algorithms for the estimation of the emotional state of the humans and speaker estimation from facial expressions and head movement	M18	Measurement of turn-taking efficiency at Demo 2	Measurement of turn-taking efficiency at Demo 3	Measure added efficiency (speed, effort, error rate) of collaborative extraction
9	Human-leg and torso tracker using 3D information from the navigation camera	M17	Measurement of tracking efficiency at Demo 1	Measurement of tracking efficiency at Demo 2	Measure added efficiency (speed, effort, error rate) of collaborative tracking
10	Participant Negotiation Module, distributed reasoning/data fusion system for estimation of observations of the participants.	M17	Measurement of collaborative data sharing at Demo 1	Measurement of collaborative data sharing at Demo 2	Measure added efficiency (speed, effort, error rate) of collaborative data sharing

The report is organised as follows: First we describe the *Experimental Setup*, then we present *Results of Demo-3* based on figures collected from (multiple runs of similar) scenarios relevant to key performance indicators (KPIs).

KPIs from Demo-1 and Demo-2 are relevant here and included in Table 1.

The last section called KPI Analysis contains a description of how the project has met each of the technical KPI agreed with the project group during the negotiation phase and subsequent adjustments.

EXPERIMENTAL SETUP

This section provides a short description of, in the following order, *physical space*, *robot hardware*, *robot software*, *measurements*, and *experimental procedure / run*.

Physical Space

The demonstration took place in IIIM's offices in Reykjavik within an area of approximately 3 x 6 meters. The lighting consists of built-in overhead fluorescent lights. The local Wi-Fi network provided communication between the robots and the base computers. The experimental setup for CoCoMaps Demo-3 was very similar to Demo-1 and Demo-2. Two control panels were arranged at one end of the space, with approx. 1.5 meters between them with which the robots interacted virtually (as they have no arms and hands).

Demo-3 Robot Hardware

We use two identical TurtleBot2 robots[1] identical to those in Demo-2 including the better RGB camera, used for human detection and recognition, sitting on a new custom stand that raises it higher from the robot base, to better avoid glare from the overhead fluorescent lighting. The new camera is a Logitech BRIO with a resolution of 1920 x 1080 pixels, using raw uncompressed video, sufficient for the human detection and recognition module, which requires high definition camera to support increased working distances for face recognition.

The main computer is as before an Intel NUC, placed onto each TurtleBot structure.[2]

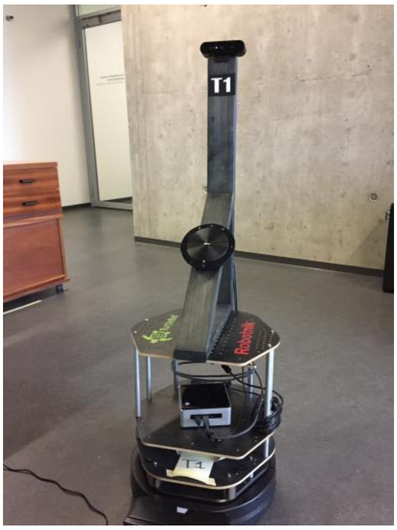

Figure 2.

TurtleBot 2 with the *Kobuki* base, including an *Astra Orbbec 3D* depth camera and an *Intel NUC* control computer. The *Logitech BRIO* USB camera on a stand, which also includes the *Jabra Speak* integrated microphone and speaker.

[1] TurtleBot 2 is an open-source hardware project built on the mobile Kobuki (http://kobuki.yujinrobot.com/wiki/online-user-guide/) base. The base supplies power for the entire system, has a motor to move through the surroundings as well as sensors used in navigation. TurtleBot 2 comes with setup for a 3D depth camera that can be used for mapping and localization. The Kobuki base uses a standard 12 V brushed DC motor. The batteries are Lithium-Ion 14.8V 4400 mAh, 4S2P configuration. Additional sensors used in navigation are a 3-Axis digital gyroscope from STMicroelectonics, part name L3G4200D, with a measurement range ±250 deg/s. Additionally the base comes with 3 bumper sensors, left, center, right. The complete structure is cylindrical with a diameter of 354 mm and height, from floor to top of the structure 420 mm. The Kobuki base has ground clearance of 15 mm. The combined weight of the base and structure is 6.3 kg, without the computer, USB camera and other additional peripherals. See http://www.turtlebot.com/turtlebot2/.

[2] The specific NUC used is the NUC5i7RYH. It has an Intel Core i7 processor, uses 8GB DDR3 memory, an integrated graphics card and Wi-Fi. Further information: https://ark.intel.com/products/87570/Intel-NUC-Kit-NUC5i7RYH.

For navigation, mapping and localizing a 3D depth camera, Astra Orbbec, is placed in the centre platform of the TurtleBot structure. The camera has a range of 0.6-8.0 m with a maximum depth image size 640x480 at 30 fps.[3]

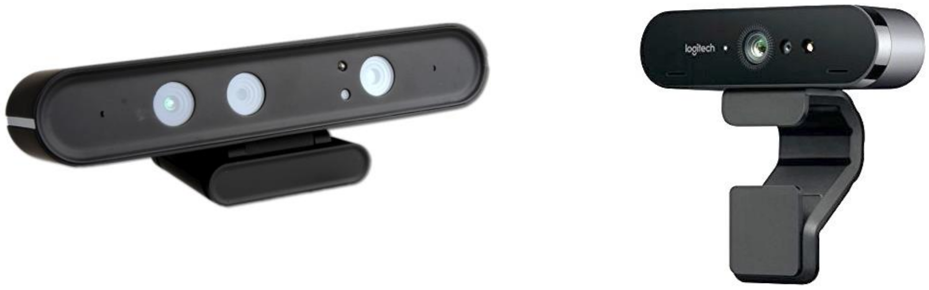

Figure 3.
Left: The Orbbec Astra 3D depth camera, mounted on the center platform of the turtlebot. *Right:* The new Logitec BRIO camera mounted on the top platform of the TurtleBots.

DEMO-3 Robot Software & Architecture

As in Demo-1 and Demo-2, the robots run identical software, but maintain a separate local current state and have separate IDs. As before, each robot runs a Psyclone 2 system which contains a number of modules and catalogs. Underneath Psyclone the ROS system interfaces with the actual hardware sensors and motors.[4]

The components running in the Psyclone system relevant for Demo-3 are listed in Table 2 below. Catalogs can be seen as containers and arbitrators of data while modules are the processors, detectors and decision makers.

The robots communicate via the CCMCatalog (as in Demo-1 and Demo-2). At this stage the CCMCatalog is used to share information on humans that have been detected. All robot decisions are made independently by each robot – the CCMCatalog acting as a centralised storage for observations, providing a virtual channel for the robots to negotiate with each other about sub-tasks including where a human is located, where each should navigate next to ensure best observation coverage, and their own position in the scene.

To update the CCMCatalog each robot has a separate CCMCollector module that collects relevant data and communicates with the CCMCatalog. All observations of humans detected in the scene are continuously updated to the CCMCatalog by the CCMCollector. Each observation is tagged with metadata: (a) who made the observation, (b) when, (c) where and (d) the confidence of the correctness of the observation. Each robot can query the CCMCatalog for all such metadata.

[3] See https://orbbec3d.com/product-astra/.
[4] More information: http://cmlabs.com/products

Table 2.
Key software components used in Demo-2 and Demo-3.

COMPONENT	ROLE
CCMMaster Type: CCMCatalog	The central CCMCatalog which holds all the shared information in the whole system. Only one of these exists for each full system and each robot connects to this via the network.
DemoRecording Type: ReplayCatalog	Catalog that makes a recording of all the relevant messages in the system for later analysis of time and resources spent, timing of detections and decisions, etc. It takes no active part in the demo itself.
MessageDataCatalog Type: MessageDataCatalog	This catalog stores messages and their associated data for human viewing and debugging the system. It takes no active part in the demo itself.
PositionCollector1 Type: CCMCollector	This catalog collects local information about object (both robots and humans) and loads the information into the shared CCMCatalog. It will also allow querying based on time and space and allow the robots to negotiate about the position of objects in the scene.
RobotStatus Type: Module	The ROS system interface. It uses ROS to gather data from the robot sensors including the cameras and allows other modules to send commands to the robot such as navigation and turning.
RobotSelf Type: Module	This module analyses all the data gathered from the robot itself and converts this into the Psyclone data architecture. It also keeps the CCMCatalog up to date with the latest state, position, etc.
RobotNavigation Type: Module	Performs the search pattern negotiation via the CCMCatalog to agree with the other robots on where it should go next. It also allows a human operator to override the current navigation pattern and pauses the search pattern when the robot is currently tracking a human in the scene.
FaceRecognition Type: Module	Module that receives the video stream from the USB camera on the robot and analyses it for faces. For every face found it performs an identification as well as facial expression analysis.
HumanDetection Type: Module	This module keeps track of the faces and humans detected in the scene and from a variety of data in the system it attempt to match the face with a body and/or legs and from this and its own position and orientation will calculate the actual scene location of the human.
FaceFinder Type: Module	Module for finding faces in each video frame.
RobotSelf Type: Module	Module that collects all data relevant to the robot, including its position, orientation, identity, and current role.
SpeakerOutput Type: Module	Receives text to be spoken and plays it; manages pausing audio (during hesitations), flushing speech output buffer.
RobotSpeechMonitor Type: Module	Keeps track of which robot is speaking when.
StopSpeakingDetector Type: Module	Special high-speed detector for managing stops and starts during multi-party dialogue.

SpeechRecogniser Type: Module	This is the front-end module to the Nuance speech recognizer that interfaces with the Psyclone system. Receives recognition packets from Nuance and posts as Psyclone messages.
OverlapDetector Type: Module	Dedicated module for detecting when overlaps in speech occur. Used by robots to flush speech recognition buffer to clear misrecognitions (guaranteed to be faulty during overlapping speech).
TaskDialogManager Type: Module	Manages the interaction with humans and tracks the state of tasks that the robots are engaged in. The TDM handles context-dependent interpretation of actions and speech acts (commands, requests, etc.), manages task progress and robot task division of labor. Manages task and sub-task navigation using a task tree.
MeaningExtractor Type: Module	Responsible for turning the user's behavior into context-sensitive responses. Receives text (and, in future, gestures, facial expressions, and more), parses it, maps it into reified 'meaning structures' that are used to compose response (real-world action and/or dialog act).
RoleNegotiator Type: Module	Responsible for negotiating either shared or exclusive roles for the robots.
TaskNegotiator Type: Module	Responsible for negotiating which of the robots should carry out a task, based on their current roles and other parameters.

Table 3.
New software components important in Demo-3.

InterruptionDetector Type: Module	Detects interruptions. Used by Turn-taking module, TaskDialogManager, and others.
DigitsViaSpeech Type: Module	Handles digit-from-text conversion.
PitchTrackerDetector Type: Module	Helps the system estimate which human is currently speaking.
HumanDetector Type: Module	Processes facial and 3D data for speaker estimation, emotions and torso and leg detection
Other Components	Numerous other system components have been developed that are fundamental (navigation, motor control, etc.) and not detailed here for brevity sake or because they are not essential for Demo-3.

MEASUREMENTS & METHODOLOGY

In human-robot interaction it is ultimately the whole overall experience that matters to the end-user. For multi-turn interactions like those demonstrated in Demo-2 and Demo-3 the overall experience is dictated by the performance and coherent operation and interaction of (most or all of) the system’s sub-components.

In Demo-1 the key development target was the ability of robots to interact with the real world.

In Demo-2 we build on this and add the ability to interact in groups (two robots, one human) using language.

In Demo-3 we build on the former demos and add the ability to interact in a 4-way interaction (two robots, two humans) through real-time interaction, face-to-face dialogue, and dynamic task assignment and information extraction, where the interaction is driven by the human, putting stronger requirements on the robots for resolving missing information through the interaction.

A typical scenario involves the robots detecting a human and asking what he/she wants them to do, and the human then informing them of a named task that they do not have much detail about in their knowledge base. The robots proceed to perform a mixed-action and dialogue interaction task where parts of the task require dialogue and others require a second human to provide information (e.g. providing the PIN number for a power-down sequence).

We run similar scenarios several times during data collection to get multiple measures for comparable contexts.

Variables & Measurements

We measured a number of variables over a series of similar scenarios. Here is an account of these, broken down by measurement type. This section explains the methodology and measurement types; results on these measurements are reported in the Results section below.

Table 4.
Measurement types used in Demo-3.

Measurement Name	**Measurement of ...**	**Measurement Method**
Speed	Average of internal processing speed (architecture).	Time difference between event start and timestamp of success message. Using messages produced by relevant modules and recorded in CoCoMaps catalogs. Based on a minimum of 10 trials.
SD	Standard deviation	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$
Min	Lowest value recorded in the trials.	
Max	Highest value recorded in the trials.	

Table 5.
Variables measured in Demo-1, Demo-2 and Demo-3.

Human Detected	The time it takes a robot to know there is a human in the scene.	Wall-clock time: Timestamp (msec) of “human detected” posting minus the timestamp marking when the human entered a robot’s visual image (ground truth - timestamp generated manually by a human observer).
Person Identified	The time it takes a robot to find the identity of a person that has been detected as a human.	Wall-clock time: Interval (in msecs) between the time a human is detected until a robot correctly posts his/her identity.
Human Identified (collaborative)	The time it takes two robots in collaboration to find the identity of a person that has been detected as a human.	Wall-clock time: With both robots present, measured from the time a human enters either robot’s camera frame (timestamp generated manually by a human observer), to the time the person's identity is logged in the shared data structure (CCMCatalog).
Human Leaves	The time it takes a robot to record that a human identified as such has left its current visual frame.	Wall-clock time: Measured from the time the human leaves the scene (ground truth) until either robot posts "human left".
Speech-to-Text	Correctness of transforming audio stream to the correct words.	Percent correct of words for sentences in Demo-2 and Demo-3.
Turn-taking	The ability and time taken by a communicating robot to detect turn-taking events.	Wall-clock time: The time between a human giving turn, based on microphone signal, and the internal turn-taking state machine posting a message to the whole system to that effect.
Role Negotiation	The ability and time taken by two robots to negotiate which role each robot should assume.	Wall-clock time: From the time a human is found until both robots have confirmed their role to each other.
Task Negotiation	The ability and time taken by two robots to negotiate which robot should carry out the task.	The ability and time taken by one robot to produce the audio output stream corresponding to the speech output decision.

Human Detected

Measurements for how well humans are detected relies on the chain HumanEnters → FaceFound. This covers the initial detection of the human; tracking is done once the human has been identified. The detection process starts when a face is detected in the 2D USB camera which kicks off a series of modules to analyse the face and to search for the matching detection in the 3D depth sensor image, either the face itself, the corresponding torso or the legs. If successful the position and orientation of the robot can now be used to pinpoint the actual location of the human in the scene, both for detections and for subsequent tracking.

Person Identified

We use the same labels as for the HumanDetected part of the chain (see above), with the message type HumanAppearedSelf, which is posted by the HumanDetection module, once the location in 3D has been determined using the body and leg detector. The identification is done using facial recognition and the top three matches of enrolled faces are presented as possible identities. Subsequent sightings by the same or other robots may update the identity if a higher confidence level is achieved.

Human Identified (Collaborative)

Once a human has been detected by one robot this robot will notify the other one via the CCMCatalog. This uses the same negotiation mechanism as regular observations. (Details on the negotiation process are found in report Final Implementation of the CCM Architecture, deliverable T8.D2).

Human Leaves

To measure the robots' ability to detect when a human has left we search through the logs for the HumanLeft message, which is posted by the vision system, take note of the timestamp (clocks are synchronized across the CoCoMaps architecture) and subtract from this the timestamp of a manually marked signal in the log files for when the human actually left the image. The HumanLeft message is posted every time any human leaves the scene and contains the identity of the person. When all humans have left a separate AllHumansLeft message is posted.

Speech-to-Text

Speech-to-text is the transformation of audio signal to words. To get a baseline for dialogue understanding (see *Dialogue Understanding* below) this is an important measure because, since understanding relies heavily (but not only) on the speech output of the humans, the quality of the transformation places a ceiling on how well dialogue understanding can work.

We tested 30 sentences similar to those that are used in a typical interaction in Demo-3 and measured the number of words that were correctly transcribed to text. The speaker was a native speaker of French. The Nuance speech recognizer relies on an American pronunciation library and language model, which is not optimal for users with a foreign accent.[5]

Turn-Taking Smoothness

To evaluate the smoothness of the interaction, one measure is the quality of the unfolding turn-taking. In Demo-3 the robot always has something to respond to when the human gives the turn, and the immediacy of taking the turn is a measure of smoothness.

Tests were conducted with multiple occurrences of the human speaker giving turn to the robot, to measure how accurately the robot does take the turn.

Instances of human giving turn are symbolized in the pub-sub system by messages of type "OtherGivesTurn", produced from lower-level signals including speech and vision. If properly detected by the robot, and decided to act upon, the robot outputs the message type "IAcceptTurn". In order to measure the ability of the robot to react accordingly to human reactions, we split the dataset, looking at the succession of events that predated a

[5] In spite of repeated attempts at getting different language models for the speech recognizer, Nuance was not able to fulfill this promise according to the description and spec for their recognizer. Since much of the quality of the interaction hangs on the speech recognition working well, switching to a different speech recognizer is therefore high on the priority list for low-hanging fruit for improving the system.

"IAcceptTurn" event, using system timestamps for estimating latency between the relevant messages.

We consider that any cue of type "OtherGivesTurn" that has *not* been acted upon within 2.5 sec is lost, meaning the robot has failed to take turn. In addition, if several cues of "OtherGivesTurn" are given within the three minutes before the robot decides to take turn, all but one of these cues are effectively wasted. This enables us to measure the average "wasted time" in our turn-taking system.

Role Negotiation

In the light of the collaborative nature of the projects, the two robots assign themselves roles to perform the task at hand more efficiently. We have a 'communicator' role whose responsibility is to handle interaction and communication with human partners, while the role of a 'controller' is to act upon the information received by the communicator.

Roles can be assigned manually or assigned automatically based on the nature of the role (exclusive or shared) and the task. The task in Demo-3 required the robots to autonomously decide the roles they should assume. When a human is detected by either robot, the robots will attempt a facial recognition to identify one of their known partners. When a known human has been identified, the requirement for changing the robots' roles is met, after which role negotiation is carried out through the CCMCatalog. The robot that first recognizes the human automatically assumes the role of *communicator* and proposes this via the CCMCatalog; when the other accepts it becomes the *controller*, whose task is to execute actions in the task environment.

Further details on the role negotiation process are found in report Final Implementation of the CCM Architecture, deliverable T8.D2).

Task Negotiation

Tasks can be assigned manually or assigned automatically based on the nature of the task (exclusive or shared). In Demo-3 the robots did not know which task they would be assigned and thus had to consult a human for what to do and which role to assume. Once one robot detects the requirement for executing a task the new task will be proposed and negotiated between the two robots via the CCMCatalog. Further details on the task negotiation process are found in report Final Implementation of the CCM Architecture, deliverable T8.D2).

Table 6.
Overview of new measurements used in Demo-3. The higher each of these are, the less artificial – i.e. more natural – the interaction is.

Measurement	Estimation of ...	Measurement Method
Facing-Direction	Accuracy of the direction that a face is turned.	Comparison of head-direction angle (ground truth) to estimated output using video signal.
Emotional Reading	The ability and time taken by a communicating robot to read human emotional facial expressions.	Percentage correct emotion classification for "happy" (smiling face) over classification of "sad" (frowning face) given a set threshold based on a comparison of the two. Since "sad" is not a category in the system it serves as comparison.
Who is Speaking	Who speaks/spoke when	Event-triggered analysis of mouth movements in video images.
Dialog Understanding	The ability of the robots to say / do the right thing during the dialog	Average % correctly/successfully executed turns for an interaction scenario

Facing-Direction

The Facing-Direction variable represents the angle of a human face relative to the robot's camera. This is estimated whenever a human is detected in the image. The evaluation of the quality of this estimate was done by comparing the angle reported to ground truth for the angles 0º and +/- 30º on 20 measurements.

Emotion Reading

Experimental protocol for measuring emotion was the following: Human subject stands in front of robot doing "happy faces" (smiling) and "sad faces" (frowning) numerous times (N=10). "Sad" is not tracked separately in this system and can thus serve as a comparison to estimate and threshold the differential in scores for the smiling expressions.

Who is Speaking

While there is certainly room for improvement in this measure, it shows a better-than-random ability to estimate who is speaking, which when combined with other signals being recorded in the system can help the interpretation process produce the right action for any turn. The process consist of a temporal analysis of the mouth position and how open it is, done over the last few seconds of data. The process also supplies a confidence value which can be used by the TDM to decide amongst the available data who is actually the person speaking. Supplementing this measure with other methods for estimating who has "truly" has the floor (turn) in the dialogue, including analysing the signal from the microphones for estimating the direction the sound is coming from, would significantly improve this measure, quite possibly to a level that works for the vast majority of cooperative dialogue scenarios.

Dialog 'Understanding' [6]

Understanding dialog – i.e. interpreting the context and utterances in a way that results in pragmatically acceptable ("correct") action – is a high-level measure of CoCoMaps robot interaction performance that is effected by all levels of the system, from the hardware (quality of the microphone, acoustics of the room) and software (voice recognition software, text analysis algorithms) used, as well as the interpretation of the circumstances, control of the interaction and control of the robot body.

To evaluate how well the robot answers to human cues and how closely it follows human instruction we use a measure of action on part of the robot - how sensible a robot's action is in light of the current state of the dialogue. This measure is 'quantized' by the system's turns, which (in its simplest form) is the notion of who is accepted by both (or all) conversents to be in control of "the floor". In a two-party dialogue, "correctly taking the turn" means that, once a speaker has finished their utterance, the other speaker realizes it is their turn to speak and will then "take the turn", based on what happened so far in the dialogue. In the context of Demo-3, the robot always has something to respond to when the human gives the turn, and should always take the turn and respond to it.

The percentage of such turns wherein the robot "does the right thing" is therefore a good measure of the dialog 'understanding'.

We measure the dialog 'understanding' of the robots as the percentage of turns that result in acceptable and/or correct actions – i.e. "pragmatically successful single turns". Note that this is different from both measures of the speed or smoothness of turns and the correctness of the interpretation of a human's utterances because it takes the aims and goals of a particular interaction into consideration.

[6] While not true understanding as humans are capable of, CoCoMaps enables robots to do what could be called a "pragmatic interpretation of circumstances".

Experimental Execution

The demo consists of the following: Two idle robots in the aforementioned 3x6 meter area populated by two (virtual) control panels.[7] Whenever a human enters a scene they request the human's help for performing a sequential task involving one of the two panels. The robots do not know the steps need to perform the task, and need to extract this information from two humans via natural language. This scenario was repeated several times to produce reliable measurements for each of the target variables on the relevant dimensions, as reported in the table below.

During each run of the task the robots collaborate via the CCMCatalog to share information about humans and to negotiate roles when a task has been identified. If no human is present each robot follows the negotiated search path, as in Demo-1. When a human is observed the robots request assistance with a task they know how to perform.

To ensure that all measurements were accurate and to fix any anomalies in the experimental setup, several runs of the scenario were performed. Each run lasted approximately 10 minutes.

[7] The panels are displayed on a screen with which the robots interact via wireless messages.

RESULTS

Demo-3 data shows that the expansion of the system has not decreased reliability of its operation; as before the robots run hours at a time. While there is clear room for improvement on many measurements, it also shows that target functions perform numerically in the right ballpark.

The main results are summarized in Tables 7, 8, 9 and 10 below; Tables 7, 8 and 9 include previous measurements for comparison; Table 10 presents measurements new for Demo-3.

Table 7.
Summary of Demo-3 results for repeated measurements of Demo-1 and Demo-2. Numbers (in parenthesis) from Demo-1 (middle of cells) and Demo-2 (bottom of cells) are included for comparison.

EVENT	**Success rate** (%)	**Speed** (msec)	**SD**	**Min**	**Max**
Human Detected Interval between timestamp of “human detected” posting minus the timestamp marking when the human enters the area where the robots can detect humans	89 (78) (35)	895 (870) (2978)	780	50	2300
Person Identified Interval between timestamp of “human identified” message minus the timestamp of the “human detected” message	84 (81) (25)	8100 (8340) (3556)	6990	1520	14250
Person Identified: Collab. Interval between timestamp when the person’s identity is stored in the CCMCatalog minus the timestamp of “human detected” message	84 (81) (---)	1590 (1680) (---)	800	420	2670
Human Leaves Measured from the time the human leaves the scene (ground truth) until either robot posts message "human left"	100 (100) (80)	840 (630) (5181)	388	960	9980

The modules in charge of performing these functions are largely identical to the ones used in Demo-2 and the measurements confirm that we are seeing minor statistical variations as expected.

Table 8.
Summary of Demo-3 results for repeated measurements of Demo-2. Numbers (in parenthesis) from Demo-2 are included for comparison.

EVENT	**Success rate** (%)	**Speed Ave.** (msecs)	**SD**	**Useful time** (msecs)	**SD**	**Wasted effort**	**SD**
Turn-Taking Smoothness % turns with no overlaps and <2,5 sec pauses between turns	98 (97)	2240 (2120)	322 (362)	190 (182)	910 (830)	312 (492)	299 (363)
Role Negotiation Time and effort measurement from one robot deciding that a role needs to be assigned until the negotiation has been completed.	100 (100)	0.44 (0.42)	0.12 (0.087)	0.55 (0.42)	-	0 (0)	-
Task Negotiation Time and effort from one robot deciding that a task needs to be carried until the negotiation has been completed about which robot has accepted the task.	100 (100)	0.068 (0.068)	0.048 (0.038)	0.41 (0.42)	-	0 (0)	-

Turn-Taking Smoothness

We measure the efficiency of the Turn-Taking system by the average time difference between the reception of the human speech input and the decision taken by the system to act upon it.

- Proportion of turn-taking events one or more wasted cues: 28%
- Proportion of failed turn-taking events (robot did not take turn): 22.4%
- Success rate (robot took turn): 77.6%

Role Negotiation

As the numbers indicate, Role Negotiation (which robot is 'communicator' and which one is 'task executor') is a fast and seemingly bug-free operation. The negotiation mechanism, and their supporting processes, operate very reliably and efficiently.

Task Negotiation

The same can be said of task negotiation as of role negotiation, which in large part relies on the same mechanisms (but not entirely). Negotiation of tasks happens internally to the robots, deciding which one has to accomplish which given task.

Table 9.
Summary of Demo-3 results for repeated measurements of Demo-2. Numbers (in parenthesis) from Demo-2 are included for comparison.

Measure	**Ave % Correct**	**Description of Measure**
Speech-to-Text % correctly transcribed words from speech	54 (66)	Average of words that are transcribed correctly by the speech recognizer (Nuance) during dialogue.

Speech-to-Text

In comparison to Demo-2, this demonstration had a much larger vocabulary and a significant decrease was expected. A drop of 12 percentage points is better than expected and can be explained by the fact that the humans used longer sentences which increased the efficiency of the semantic analyser in the Nuance product.

Table 10.
Summary of new Demo-3 measurement results: Facing-Direction, Emotional Reading and Dialog Understanding.

Measure	**Ave % Correct**	**Description of Measure**
Facing-Direction The measurement of correctness versus false estimates.	~92	Based on largest mean absolute error (MAE=4.4) as over the full practical (60º) angle for which the measurement is possible
Emotional Reading The measurement of correctness versus false estimates.	73	The ability by a communicating robot to read human emotional facial expressions.
Speaker Estimation The measurement of correctness versus false estimates.	59	The ability by a communicating robot to estimate whether a human face is the current speaker.
Dialog "Understanding" % turns resulting in correct robot action / event	66	The ability of the robots to say / do the right thing during the dialog

Facing-Direction

For small angles the precision is greater. The robot has an easier time correctly detecting humans - and the head angle - when facing west (meaning, the robot was on the right from the direction I was looking at) than when facing east. However, when detected east, the measure is slightly closer to the actual angle.

Emotional Reading

For the measurement of accuracy we used one of the categories, the "happy face". We ran 22 measurements of a smiling face. Out of these 22, in 16 cases the dominant emotion detected was "happiness" => the "Success rate" (rate of true positives) is 73%. In 4 cases, the dominant emotion detected was "surprise" and in the two remaining the emotion detected was "neutral" (which means either 18% or 27% error rate, depending on whether we consider "neutral" an error or just misdetection).

Scores are given from 0 to 100. The average "happiness" score was 54.1 with a standard deviation of 17.7.

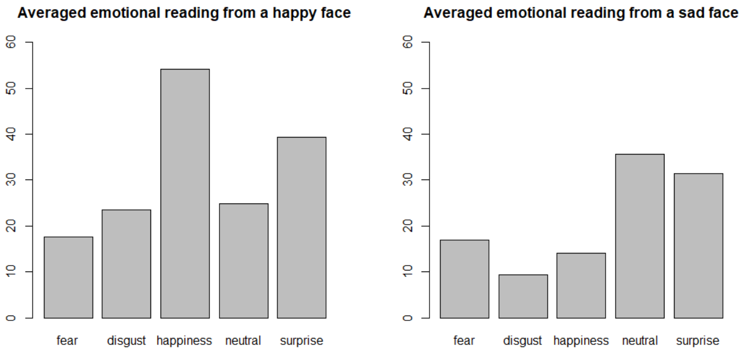

Figure 4. Average readings from the Emotion trial runs.

Who is Speaking

In multi-party dialogue it is important to know who has the floor ("turn"). A good measure of this (albeit by no means the only measure in more complex human dialogue) is who is speaking. While seemingly simple, even measuring reliably who is speaking in a multi-party conversation may be a challenge. The method we have chosen seems to work reasonably well, while leaving some room for improvement.

Dialog Understanding

Using the Nuance system we achieve a precision of ~66% of correct recognition of sentences. This can undoubtedly be improved, in spite of a fairly noisy environment, by changing to a more appropriate language model (the current one is for an American accent). So far we have improved the understanding of dialog by 12% from what it was originally, by special handling of the most common misrecognitions, especially words that sound the same way but have different meanings. This has resulted in an average of 66% understood turns. Since dialog understanding relies not only on speech recognition but a host of other data and processes, this number should in fact be somewhat lower than the speech-to-text transcription number. The fact that it isn’t is in large part due to our ability to interpret interaction contextually, via the chosen design of the Task-Dialog Module (TDM) and Meaning Extractors (MEx).

KPI ANALYSIS

KPI 1: Ability of current state of the art running on one computer

In Demo-3 CoCoMaps runs on three computers - on each of the robot's computers and on a third offboard computer. Each robot computer runs the full system which includes the state of the art for turn-taking and the cognitive map. This KPI has been met.

KPI 2: Ability of real-world robot-robot interaction using new collaborative CMArch

In Demo-3 and previous demonstrations the robots demonstrate the ability to share and query data, as well as negotiate roles and tasks in near real-time, with (near) 100% success rate for several runs, and within very acceptable time frames (see Tables 7-10). This KPI has been met.

KPI 3: Ability of real-world multi-robot-human interaction using collaborative CMA and speech

Demo-3 has shown that our system is able to allow multiple robots to collaborate successfully both between themselves using the CCM architecture and with multiple humans via speech and natural dialogue (see Tables 8-10). This KPI has been met.

KPI 4: Efficiency of collaborative detection of humans

Collaborative detections of humans were proven in Demo-1 and further refined and significantly improved in Demo-2 and Demo-3. The efficiency measurements are provided in Table 7 and show that two robots can more effectively detect humans when collaborating on the task. This KPI has been met.

KPI 5: Efficiency of collaborative tracking of humans

Collaborative tracking of humans was proven in Demo-1 and further refined and significantly improved in Demo-2 and Demo-3. The efficiency measurements were provided in Table 7 and show that two robots can more effectively track humans when collaborating on the task. This KPI has been met.

KPI 6: Efficiency of collaborative information extraction through dialogue

Collaborative information extraction from humans was proven in Demo-2 and further refined and improved in Demo-3. The efficiency measurements were provided in Table 10 and show that the robots can both extract information and perform a remote task at the same time. This KPI has been met.

KPI 7: Efficiency of collaborative task extraction through dialogue

Collaborative task extraction from humans was proven in Demo-3. The efficiency measurements were provided in Table 10 and show that the robots can both extract the task and the required information as well as perform a remote task at the same time. This KPI has been met.

KPI 8: Real-time algorithms for the estimation of the emotional state of the humans and speaker estimation from facial expressions and head movement

The algorithms for estimating emotional state of humans and speaker estimation were used and shown in Demo-3. The efficiency measurements for both were provided in Table 10 and show that the system is able to use visual analysis of the humans' faces to estimate emotions, head movement and speaking activity. This KPI has been met.

KPI 9: Human-leg and torso tracker using 3D information from the navigation camera

The algorithms for detecting legs and torso were used in the HumanDetector module, used in Demo-3 to detect and track the 3D position of the human. The efficiency measurements for this were included in the Human Detection entry in Table 7 and show that the robots are able to estimate the position in the room by using 3D depth information to find either the torso or the legs of the person. This KPI has been met.

KPI 10: Participant Negotiation Module, distributed reasoning/data fusion system for estimation of observations of the participants.

The module for negotiating observations of humans were demonstrated in Demo-3 and the functionality was used every time the two robots shared information about their observations. Specifically, this was measured in the **Person Identified: Collab.** entry in Table 7. This shows that the robots are able to discuss and negotiate both observations, roles and tasks with other robots via the CCMCatalog. This KPI has been met.