2013 IEEE/RSJ International Conference on
Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Joint Action Understanding improves Robot-to-Human Object Handover

Elena Corina Grigore[1], Kerstin Eder[1], Anthony G. Pipe[2], Chris Melhuish[2] and Ute Leonards[3]

Abstract— **The development of trustworthy human-assistive robots is a challenge that goes beyond the traditional boundaries of engineering. Essential components of trustworthiness are safety, predictability and usefulness. In this paper we demonstrate that the integration of joint action understanding from human-human interaction into the human-robot context can significantly improve the success rate of robot-to-human object handover tasks. We take a two layer approach. The first layer handles the physical aspects of the handover. The robot's decision to release the object is informed by a Hidden Markov Model that estimates the state of the handover. Inspired by human-human handover observations, we then introduce a higher-level cognitive layer that models behaviour characteristic for a human user in a handover situation. In particular, we focus on the inclusion of eye gaze / head orientation into the robot's decision making. Our results demonstrate that by integrating these non-verbal cues the success rate of robot-to-human handovers can be significantly improved, resulting in a more robust and therefore safer system.**

I. Introduction

Human-assistive robots are machines designed to improve the quality of our lives by helping us to achieve tasks. Such robots act within the personal space of a human, including human-robot shared manipulation of objects and even direct physical contact. While the actions of a specific task a robot performs remain largely the same, every single execution of this task will be slightly different in detail. This is due to the constant change of the exact situation in which everyday tasks are performed in human lives and the innate variability in human performance. Thus, the robot is required to constantly adapt its behaviour to different situations.

To be genuinely useful, some robots may need to be powerful (e.g. to support the weight of a human) and therefore are potentially dangerous. This raises concerns about whether human assistive robots can be trusted with respect to human safety. To ensure that personal robots entering widespread use do not pose a serious risk for humans interacting with them, we here suggest that robots need to be able to take into account what the human in their close proximity is doing when planning their own actions.

Imagine, say, a basic scenario in which you wanted a robot to pass you a cup of hot tea. In contrast to other days, you are just about to write an important manuscript and would like to continue working while receiving your cup. How would you ensure that the robot knew when you were ready to receive the cup without disturbing you through asking? How would the robot pass you the cup without harming you by spilling tea over your hand or dropping the cup because you were not grabbing it in time? And how would the robot be able to take into account the dynamics of your own hand and arm movements while you reach out for the cup with only a brief glance? In more general terms, how can the decisions and actions taken by an autonomous personal robot working within the personal space of a human be made safe, reliable and predictable, given the large variability of human actions for even such a simple task?

How to design robots that are entrusted to act and interact in an environment constantly modified through the breadth of human action is a research challenge that requires expertise beyond the classic fields of robotics. In particular, it is critically important to utilize the latest findings from joint action research in the context of human-human interaction [1] when developing human-assistive robots. Increased understanding of how humans interact with each other is expected to lead to more informed engineering decisions during the design of human-assistive robots; this in turn results in more robust and therefore safer Human-Robot Interaction (HRI).

Humans constantly move their eyes and orient their heads toward objects of interest, signals that can be evaluated by others (or a robot) to obtain information about their counterpart's focus of attention and thus indirectly their engagement in the interaction [2]. In this paper we investigate whether safety in a basic object handover task from a robot to a human can be increased, if the robot is able to take into account non-verbal cues about its human co-worker's actions. Our intention is to make the robot-to-human handover task as natural as possible to increase safety. We use the success rate of the handover as our metric for safety in this context.

We developed a two-layer system. First, we focused on the physical aspects of the handover. Based on a model that captures the different stages during a handover in a set of distinct states, a learning algorithm was used to assess the situation at any stage during the interaction and to estimate its state. To release the object the system was required to be in a state in which it was considered safe to do so. Information for the state estimations was obtained directly from the robot's arm and hand holding the object. This basic layer provided an adequate estimate of the handover dynamics. Sensing when the human applies force is, however, not sufficient for *safe* object release, which requires both mechanical pressure *and joint attention*.

In a second step, we therefore superimposed a higher-level

[1]E.C. Grigore and K. Eder are with the Department of Computer Science, University of Bristol, Merchant Venturers Building, Bristol, BS8 1UB, GB {eg9542,Kerstin.Eder}@bristol.ac.uk

[2]A.G. Pipe and C. Melhuish are with the Bristol Robotics Laboratory, T Block, UWE, Frenchay Campus, Bristol BS16 1QY, GB {Tony.Pipe, Chris.Melhuish}@brl.ac.uk

[3]U. Leonards is with the School of Experimental Psychology, University of Bristol, 12A Priory Road, Bristol, BS8 1TU, GB Ute.Leonards@bristol.ac.uk

978-1-4673-6357-0/13/$31.00 ©2013 IEEE

cognitive layer that monitors the humans focus of attention during the handover task. It captures a model of actions (including the most likely order of events) characteristic for an attentive human user in a handover situation, derived from human-human handover observations. This level includes evaluation of subtle information such as human eye-head gaze shifts [3] which provide engagement indicators. The goal of including this higher-level layer is to increase safety when the most likely action sequence of normal handovers is not followed. In such cases the robot then does not release the object and either puts it back or takes further steps to ensure the user still wants it. Our experiments demonstrate that integrating these cues into the decision making process of the robot significantly increases the success rate of the handover, and therefore results in a system that is overall safer than the basic single layer version.

II. Background and Related Work

A critical pre-requisite for humans to accept robotic assistants acting in their personal space is that they are demonstrably safe, predictable and hence trustworthy [4]. We hypothesise that robotic assistants can only be trusted as safe to interact with and will only be accepted into our homes if they are designed to take human behaviour and intentions into account when planning their own actions. This hypothesis is derived from three to date quite separated fields of research: a) safety within personal robotics, b) handover tasks as an example of human-robot/human-human joint action, and c) joint attention. The state of the art in each of these fields is considered in turn below.

A. Safety

Robotic software architectures are often layered [5], [6]. The low-level layers generally deal with control systems, while high-level ones deal with the robot's knowledge, goals, and plans. While considerable effort has been invested into assurance of low-level physical safety properties in robotic controllers, the higher levels at which the decision making takes place have received very little attention to date.

Current research is focused on methods that guarantee the functional correctness of robotic systems either by design time verification or by systematic design. Examples include the application of model checking to prove that a set of safety properties is satisfied by an adaptive multi-agent control system [7]. Other work has focused on ensuring safety by construction [8]. Using a systematic component-based design, construction and verification approach can enforce safety properties by design [9]. Combinations of formal with simulation-based methods are also being developed for the verification and systematic refinement of safe robotic controllers [10], [11].

The critical issue of ensuring the high-level behavioural safety in HRI, however, remains a challenge. In [12] an initial set of safety and liveness (a.k.a. usefulness) properties has been explored as key foundation towards the trustworthiness of a human assistive robot in an object handover. This research has highlighted the need to closely integrate the human's behaviour into HRI safety considerations.

In practice, the process of confirming that a system satisfies its formal requirements (verification), is often separated from the process of confirming that the system results in the intended behaviour once it has been integrated in its target environment (validation). Establishing safety in HRI clearly calls for a tight coupling of these two processes and a detailed investigation of the human-robot interface.

Creating trust at this interface necessitates a transfer of our understanding of human-human interactions into the HRI context. This is because trust in robots, just like in humans, needs to be earned. We therefore need to first understand the properties that humans seek when establishing trust, i.e. those associated with safety and usefulness of the robot. Once established, these can be integrated into the robot's decision making, resulting in a safer, more trustworthy system.

B. The Handover Process

Two vast, mainly separate strands of literature in psychology and the cognitive neurosciences provide crucial insight into processes underlying object handover in humans: the first is on visually-guided grasping and handling of objects (see [13] for a recent review), the second investigates the mechanisms underlying observing other people perform such visually-guided actions (see [14] for a recent review). Surprisingly little is known about how these aspects are merged in human-human social motor coordination / joint action [1], [15], where one's own actions have to be precisely coordinated with those of another person. Indeed, only a few studies have been published on joint action tasks, in particular for passing an object from one person to another [16], [17], [18]. These studies concentrated on the physical (primarily manual) aspects of the handover with its temporal and spatial parameters. Issues tackled included hand grip and load forces with their temporal dynamics, arm and upper body movement trajectories of the people involved in the process, the relative location of the object between handover partners at the time of handover, and the velocity patterns of movements within the transfer. Where exactly the two partners looked during the handover task, however, had not been considered, even though both the visually-guided grasping literature and the literature on observing other people performing tasks, indicate that visual perception might be a key factor to improve the precision of the movement and the dynamics of the manual transfer. Even more importantly, as proposed by theories on joint action [15], [1], cognitive and social aspects derived from eye-head gaze direction should provide invaluable feedback for the two partners of the state of the interaction, or their partner's engagement and readiness to participate in the task.

Even though human-robot handover studies are more numerous than human-human ones, they show a similar lack of consideration for human visual feedback mechanisms, primarily concentrating on the kinematics of the hand movements leading to the handover. In both [17] and [18] human-like trajectories leading up to the handover have

been integrated into the design of the robot's movements. A comparison of these with non-human like robot movement trajectories showed that interactions were facilitated when human expectations about natural movement were considered. This work was later extended to the actual handover mode [19], [20]. More recently, the impact of the robot's posture and its gestures on the handover has been investigated [21], again confirming that some gestures (which are more human-like) were more likely to be considered by humans as handover initiation signal. In particular, certain robotic reaching gestures were interpreted by the human as a cue to take the object from the robot [22]. Again, no information was derived from evaluating where the human looked during the interaction.

C. Joint Attention — A Fundamental Component for Joint Action

Should human eye-head gaze direction be evaluated in a robot-human handover? It makes intuitive sense that ones' own eye movements should be important to guide ones' own actions, as they direct our fovea, the point of highest spatial resolution in our eye, to the location in a visual scene that is most important for the task at hand. In most natural viewing conditions, such as visually-guided motor coordination, spatial shifts of attention and saccadic eye-movements co-occur, and attention to locations other than that of the eye-movement is strongly impaired [23]. In other words, where a person looks during a grasping task, is where they are attending to and will strongly influence the successful outcome of the action. Note that larger face shifts include coordinated head and eye movements [3].

At the same time, eye-head gaze orientation can be used by an observer to deduce where the acting person is currently attending to. Moreover, it allows the derivation of information about the observed person's intentions and mental states, or whether they need us for collaborative action [2]. If we follow another person's gaze (receiving) with our own or if we gaze-cue (send information) ourselves to create a shared space of attention that could contain objects, other people or events, we convey important functions that aid the sending and receiving of social information [24], summarised under the term "joint attention". Joint attention plays a critical role in the development of social cognitive skills, interaction and communication [25], and is at the core of "Theory of Mind", i.e. the attribution of beliefs, goals, and desires to other people.

In our handover scenario we can then predict that the human's eye-head gaze direction can inform the robot about the human's current focus of attention and their readiness to engage in the interaction. Integrating knowledge about the human user through analysis of their eye-head gaze direction (i.e. the "Theory of Mind" concept), should thus increase the safety of the handover because it becomes more predictable.

III. Human-Robot Interaction Scenario

The HRI scenario includes BERT2, an upper-body humanoid robot torso, interacting with a human to hand over a drink as shown in Figure 1. BERT2 was designed to investigate complex HRI, including verbal and non-verbal communication, gaze, and pointing gestures in a real world 3D setting [26]. To focus on HRI, as opposed to the challenges often encountered when using vision systems mounted on the robot, we used the VICON motion capture (MoCap) system to detect and localise interaction objects and the human's body parts in 3D space, in particular head orientation as a proxy for human eye-head gaze direction. The system has sufficient accuracy to follow the motion of human body parts, environmental features and objects using retro-reflective markers. The computing infrastructure is upheld by YARP[1], an open-source package that minimizes the effort devoted to infrastructure-level software development by facilitating code-reuse and modularity [27]. An essential component of HRI is spoken language. BERT2 uses the CSLU Toolkit [28] Rapid Application Development (RAD) which is based on the TCL scripting language to create connections between the actions the robot takes and the spoken dialog. RAD uses the Festival speech synthesis system, and recognition is based on Sphinx-II. BERT2 relies on two databases to represent the state of the world. The Object Property Database (OPDB) stores static information of all objects present in the interaction scenario. The EgoSphere is a fast, dynamic, asynchronous store of object positions and orientations. Both can be queried by other modules.

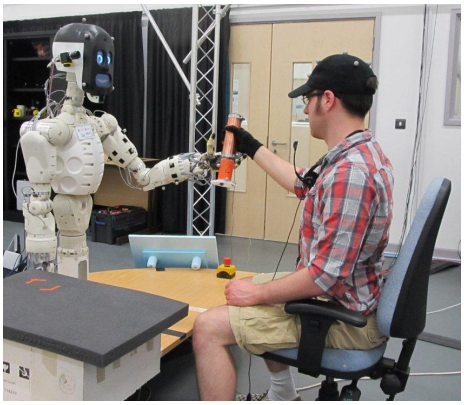

Fig. 1. BERT2 during a handover interaction.

IV. Basic HRI System — Hidden Markov Model-based State Estimation

Hidden Markov Models (HMMs) [29] are statistical models that are routinely used to model sequential, statistical processes. They have been used successfully for the analysis of temporal patterns in many areas including speech and gesture recognition. An HMM includes a set of states, state transition probabilities and also a set of observable output symbols together with their observation distribution for each state. The state in an HMM is not directly visible, but the output is. Hence, the string of symbols observed on an HMM of a process allows conclusions about the sequence of states

[1]YARP is available from `http://eris.liralab.it/yarp/`.

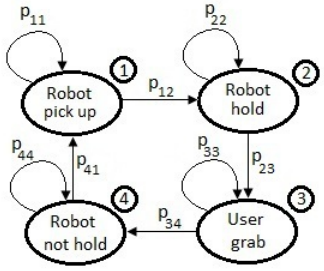

Fig. 2. HMM for handover process with non-zero probability transitions.

encountered to generate it. Formally, an HMM is a tuple $\lambda = (S, V, A, B, \pi)$ as follows:

- $S = \{s_1, \ldots, s_N\}$ is a finite set of N states. The state at time t is denoted as q_t.
- $V = \{v_1, \ldots, v_M\}$ is a finite set of M distinct observation symbols in the alphabet, corresponding to the output of the system.
- $A = \{a_{ij}\}$ is the state transition probability distribution, where $a_{ij} = P[q_t = s_j \mid q_{t-1} = s_i], 1 \leq i, j \leq N$.
- $B = \{b_j(v_k)\}$ is the observation symbol probability distribution, where $b_j(v_k) = P[v_k \text{ at } t \mid q_t = s_j], 1 \leq j \leq N, 1 \leq k \leq M$.
- $\pi = \{\pi_i\}$ is the initial state distribution, where $\pi_i = P[q_1 = s_i], 1 \leq i \leq N$.

HMMs are a natural fit our problem. The underlying state of the system (the state the robot is in with respect with the cup) is hidden and can be uncovered based on the robot's motor current and torque values (which represent the observations). Each hidden state has a probability distribution over the possible next states, which models the fact that our robot moves from one state to another with some probability based on the values observed in the current state. Moreover, there is a certain sequence of hidden states that is normally expected for the present system, which can be computed, together with the probability of their of occurrence, based on the parameters of the model.

For the basic system the HRI scenario was modelled using an HMM as depicted in Figure 2. Analysis of the handover process identified four basic states ($N = 4$): the robot is picking up the cup, the robot is holding the cup (without the user touching it), the user is grabbing the cup (joint holding of cup), and the robot is not holding the cup (has released the cup). A small number of states facilitates accurate differentiation between them and provides the robot with a clear indicator on which to base the decision to release the cup, i.e. the "user is grabbing the cup" state.

Although the model is in principle ergodic, meaning that each state can be reached from every other state, the transition probability distribution matrix A was initialized with some zero values to encourage the system to reestimate those close to zero, e.g. a transition from the robot not holding the cup state to the robot holding the cup (without going via the robot is picking up the cup state) is unlikely, but could result e.g. from erroneous transmissions of the motor current values while the robot is picking up the cup.

The initialisation of the observation symbol probability distribution matrix B is very important for a correct reestimation of the model parameters. A representative training sample of observations, one that incorporated all the usual transitions between states, was collected and partitioned into parts corresponding to the four states of the HMM. For each state the number of occurrences of each symbol was counted; the values were used to initialize B. This proved to be a sound technique in that it provided the expected results.

To choose an input signal to use for the HMM, we recorded several typical handover interactions (5 sets of 20) and analyzed both the motor current values from the robot's fingers and the torque values from the robot's arm. Based on the training data and on experimentation with different signals, including combining some signals (summing the values), the best results were obtained when using the motor current values from the robot's middle finger. This is also intuitive because the middle finger has a prominent role in every step of the motion of picking up, holding and releasing an object.

The number of distinct observation symbols per state corresponds to the number of values the input stream takes. Throughout the entire experiments we observed the maximum value of 120 and therefore $M = 120$. The initial state distribution π favours the first state (the robot picking up the cup). This biases the start of each handover, but does not influence the reestimation values (a uniform distribution provides similar results).

The reestimation technique is based on the Baum-Welch Algorithm, using formulae in [29], with a normalisation following the formulae from [30] as follows:

Forward Algorithm:

$$\hat{\alpha}_t(i) = \frac{\pi_i b_i(O_1)}{\sum_{k=1}^{N} \pi_k b_i(O_1)} \tag{1}$$

$$\hat{\alpha}_{t+1}(i) = \frac{b_i(O_{t+1}) \sum_{j=1}^{N} \hat{\alpha}_t(j)\ a_{ji}}{\sum_{k=1}^{N} b_k(O_{t+1}) \sum_{j=1}^{N} \hat{\alpha}_t(j)\ a_{jk}}, 1 \leq i \leq T \tag{2}$$

where the forward variable $\alpha_t(i) = P(O_1 O_2 \ldots O_t, q_t = s_i \mid \lambda)$ represents the probability of the partial observation sequence $O_1 O_2 \ldots O_t$ (until time t) and state s_i at time t, considering model λ.

Backward Algorithm:

$$\hat{\beta}_t(i) = \beta_t(i) \prod_{k=t+1}^{T} \eta_k, \tag{3}$$

where η_k is the normaliser

$$\hat{\beta}_t(i) = \beta_t(i) = 1 \tag{4}$$

$$\hat{\beta}_t(i) = \eta_{t+1} \sum_{j=1}^{N} \hat{\beta}_t(j)\ a_{ij}\ b_j(O_{t+1}), 1 \leq t \leq T, \tag{5}$$

where the backward variable $\beta_t(i) = P(O_{t+1} O_{t+2} \ldots O_T \mid q_t = s_i, \lambda)$ represents the probability of the partial observation sequence from $t+1$ to the end, given state s_i at time t, and the model λ.

The discovery of the most likely state sequence in a handover was implemented using the Viterbi Algorithm as

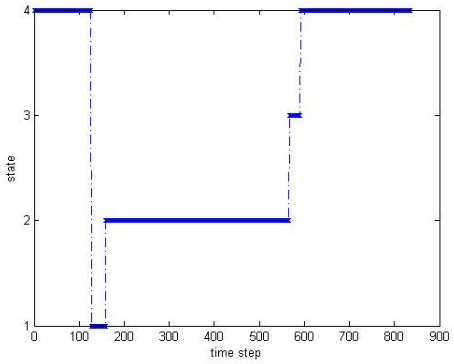

Fig. 3. Recovery of the most likely state sequence. State 1: "Robot is picking up the cup.", State 2: "Robot is holding the cup.", State 3: "Human is grabbing the cup.", State 4: "Robot is not holding the cup."

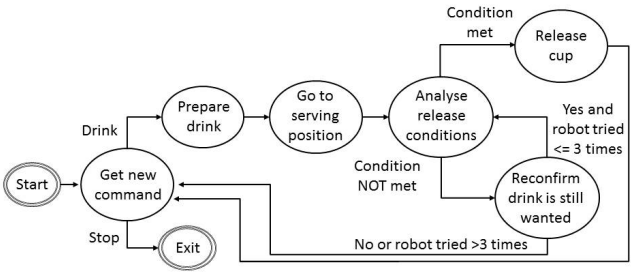

Fig. 4. State machine from the robot's perspective.

described in [29]. A result is given in Figure 3. As can be seen, the identification of State 3 is critical for the robot to decide when to release the cup. The value for the probability that the interaction is in State 3 has been determined experimentally and has proven to be a good indicator for the handover actually being in this state.

V. Extended System — User Intentions and Reactions Modelling

We have observed that, in a successful handover, the receiver performs the following sequence of actions: the user first browses the environment as their attention gets caught by different objects or people and their attention is not yet on the task at hand; the user then looks at the object, (the user might look away), the user grabs the object and the handover is completed when the robot lets go of the object. Note that the phase of the human looking away from the cup after looking at it is optional: a deeply engaged user might constantly look at the cup, another user might look at the robot's head e.g. seeking rapport, and then back at the cup. Thus, this phase can include alterations of the user's behaviour, both between trials with the same user as well as between users. The system generally detects the user looking to and away from the object several times during a handover.

The most important element, however, is that the amount of time passing between the user's first look at the cup, and the actual grab does not go above (or below) a certain threshold which was obtained experimentally. If the time is too short, this can be considered as a lack of the receiver's attention toward the object, if the time is too long, this indicates most likely that the user lost interest in the object or got distracted. When the robot detects either one of these cases, it does not consider it safe to release the cup.

As shown in Figure 4, the robot asks the user if they would like a drink. If the answer is "Yes", it "prepares" a drink (i.e. it gets the cup from a pre-defined pick-up position), moves it to a pre-defined serving position, and tells the user to take the cup. At this point, the robot starts checking the release conditions, i.e. is the user following the sequence of actions that indicates their engagement in the interaction: *user looks at cup* $\Rightarrow$ *(user looks away* $\Rightarrow$*) user touches cup*? The module that estimates the human's focus of attention works constantly in the background. Therefore, the main module can obtain, at any time, the moment when the user first looked at the cup and the moment when he or she stopped looking at it. The only value needed at this point is the difference between the current time and the time at which the user first looked at the object. This basically includes the first two actions in one interval. If the value is within the experimentally determined range, the system then checks for the user touching the cup using the basic HMM implementation. BERT2 considers it safe to release the object only when the execution of this sequence is complete.

The module which estimates the human's focus of attention works by streaming information from the VICON motion capture system regarding the position and orientation of a hat fitted with retro-reflective markers, which the user needs to wear during the interaction with BERT2. The module obtains this information from VICON and uses it to compute two vectors: one representing the direction of the human's head, and one representing the difference between the human's head and the object of interest. If the difference between these two vectors is small enough, the user's focus of attention is estimated to be on the object. This method of measuring the human's head orientation proved to be far more robust than using the gaze-tracking system and the cameras mounted on BERT2's head. The only limitation was that eye movements which are not accompanied by head movements could not be detected. In everyday tasks, however, most relevant eye-movements indicating a change of the focus of attention can be expected to be large and would thus include head movements [3]. Thus, head direction is a sufficiently accurate approximation for eye gaze in the type of physical interaction scenario (reach and grasp) we investigated, and was therefore considered sufficient as a proof of concept. For our experiments the advantages of the VICON regarding robustness outweighed this minor drawback.

To validate the HMM implementation, we also developed an alternative to check when the user is touching the cup. It uses a glove fitted with adhesive copper contacts and a long-shaped object acting as the cup in the interaction which is also fitted with copper coating, as well as with

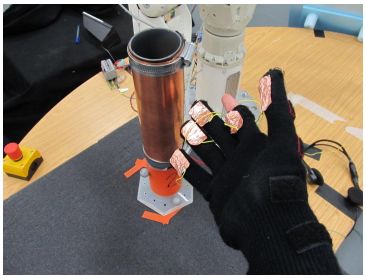

Fig. 5. Alternative setup for detecting when the user touches the cup.

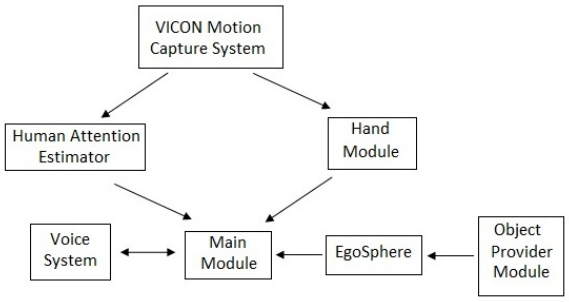

Fig. 6. Overview of the system architecture.

retro-reflective markers (for the VICON system to be able to track its position). A Phidget interface kit was programmed to signal to the main module when the user touches the cup. The set-up can be seen in Figure 5. This method precisely identifies the exact moment when the human touches the object and is therefore useful for obtaining clearly defined intervals for the actions described above.

VI. System Architecture

The architecture for the HRI system can be seen in Figure 6. The VICON Motion Capture system detects and localises the cup object and the user's head in 3D space based on retro-reflective markers. The information from the VICON software is captured by the Human Attention Estimator Module and the Hand Module. The Human Attention Estimator Module works as explained in the preceding section. The Hand Module obtains information about the position of the cup object and uses the HMM implementation in order to estimate the state of the interaction (or the alternative implementation for precise timings). The information from both these modules is fed into the Main Module, which controls BERT2's movements and communicates with the voice system, sending and receiving information to synchronize the robot's actions with the dialog. The main module also implements the User Intentions and Reactions Modelling layer, when this is in use for the extended system. The EgoSphere acts as object position and orientation storage and the Object Provider Module constantly picks up on changes to objects' position and orientation and streams the information to the EgoSphere database. The dialog between the robot and the user was based on the state machine in Figure 4 and was implemented using the state-based graphical programming environment of the RAD toolkit [28].

VII. Experimental Evaluation

The experiments required users to interact with BERT2 in a drink-serving scenario. A user was sat on a chair, in front of BERT2, with no restrictions on movements as shown in Figure 1. The participant was then asked to interact with the robot as naturally as possible, knowing that the goal of the interaction was to obtain a drink. Experiments were conducted with 17 participants. The subjects included people who were familiar with the BERT2 robotic platform as well as people who had not interacted with the robot before.

The experiments included three testing scenarios. The first one was a natural interaction scenario in which the user was asked to interact with BERT2 with the goal of getting a drink. The second involved interaction while engaged in another task: the user was asked to start counting from 4 and keep adding 7, saying the result out loudly, while trying to get the drink from the robot. This scenario is linked to the concept of cognitive load theory [31] and simulates a real-world situation in which the user is engaged in a conversation or in other tasks, while trying to obtain a drink from a robot. It helps with observing how such a task affects the behaviour of the user and how much participants trusted the robot when their attention was divided between it and another task. If the user can count successfully, this suggests that the cognitive load of the other task is not too demanding, but if the total cognitive load of both tasks exceeds the capability of the subject, the highest risk task would take priority. The third type of scenario included a surprise distraction: during some of the natural interaction scenarios, a loud noise was played just before the user was supposed to get the cup from the robot. The scenario tested what happens when a user gets distracted suddenly, by an extraneous event. The safety of the handover is particularly important in this case: if the user gets distracted to the point where he or she is not engaged in the handover process any more, the robot should be able to detect that it is not safe to release the cup.

The first two types of scenarios contained three runs each. These were intertwined with a case of surprise distraction (a larger number would not have had the intended effect, as the person is already aware a "surprise" might occur and not react to it any more). This constituted one set of experiments. Three sets of experiments were run per user: one testing the basic HMM implementation, one testing the HMM implementation extended with the model of the user's intentions, and one testing the model of the user's intentions using the glove to detect when the user was touching the cup. Participants were asked to wear the glove for all settings to minimize behaviour changes between experiments.

VIII. Results and Analysis

No significant difference was found between the two alternatives of detecting when the user touched the cup while testing the extended model. The results thus refer to these two cases as "the extended model" listing six runs for each of the first two scenarios and two runs for the surprise scenario.

Table I shows a clear improvement from the basic to the extended model. The percentage of successful handovers

TABLE I

COMPARISON BETWEEN THE BASIC MODEL AND THE EXTENDED MODEL.

System	Total number of tries	Number of successful handovers	Percentage of successful handovers	Number of times cup was not released by the robot	Percentage of times cup was not released by the robot	Number of unsuccessful handovers (cup was dropped)	Percentage of unsuccessful handovers
Basic HMM implementation	119	74	62.18	34	28.57	11	9.24
Extended with user intention model	238	180	75.63	55	23.10	3	1.26

TABLE II

TIMINGS FOR THE EXTENDED CASE.

Scenario	Minimum threshold (s)	Maximum value (s)	Average (s)
Natural interaction	0.3	5.0	2.3
Counting task	0.3	5.8	3.7
Surprise distraction	0.3	6.0	2.8

is significantly greater in the extended case than in the basic one. The number of drops is reduced to 1.26% in the extended scenario, as compared to 9.24% in the basic one, see Table I for more detailed results. This is because the simple HMM implementation only takes into account values that capture the physical aspects of the handover. It does not consider engagement in the task nor joint action expectations. Most drops occurred when the motor current values were similar to the values typically encountered in the "user is grabbing the cup" case, but the user was not actually prepared to take the object (he or she was distracted or simply hit the cup without grasping it). Our results confirm that the implementation that integrates a model of what the robot can reasonably expect a user to do, e.g. following a specific sequence of actions, can reduce false positives. The percentage of the times the cup was not released by the robot in the extended case is higher than that of the times when the cup was dropped. This is a logical consequence of the robot using a stricter policy to decide when to release the cup, thereby emphasizing safety.

In the extended case, the user's level of engagement is determined by checking that the time passing between their first look at the cup and the actual grab lies between a minimum / maximum threshold. If the value is in that range, and the user is still touching the cup, the robot completes the handover. Table II presents values for the two thresholds obtained in our experiments, as well as the average values for the three types of experiment scenarios, all expressed in seconds. Note that these values are experiment specific; they do not represent absolute values.

The experiments showed that a user needs to look at the object at least 300 milliseconds before touching the cup for the handover to proceed successfully. This minimum threshold is the same throughout scenarios because it represents the minimal amount of time which signifies the participant actually looked at the cup with the intention of being engaged in the handover process. The maximum value in the natural interaction case is 5 seconds, with an average value of 2.3 seconds. The variation occurs because users take more or less time to look at the cup and then at the robot, before touching the cup. The scenario in which the test subjects are engaged in the counting task results in an increased maximum value, as well as a larger average. This was expected because participants have to switch between tasks to complete both of them. The surprise distraction scenario affected participants differently. The average value is lower than that of the counting task, but higher than that of the natural interaction. This is because some users reacted stronger than others, i.e. inter-individual variability was larger. When the loud sound was played, some participants would turn their heads and would shift their focus of attention toward the noise, while others would simply go on with the interaction and turn around after the handover was completed. The maximum value, therefore, is rather high, which is a consequence of some users' long periods of disengagement from the handover process.

The results obtained show that by integrating a model that captures the behaviour expected from a human in a handover scenario into a robot's decision making, the rate of successful handovers between robot and human can be significantly improved compared to a setting that only considers the physical aspects of a handover without regard of the cognitive side.

IX. CONCLUSION AND FUTURE WORK

Personal robots can soon become an important part of people's lives, helping them cope with various situations and performing a wide range of tasks for them. Recent advances in control engineering and robotics have enabled robots to perform increasingly complex tasks autonomously. What remains is for humans to gain trust in the resulting intelligent systems. A pre-requisite for the acceptance of personal robots is that they are demonstrably safe, predictable and therefore trustworthy social interaction partners.

The novel contribution of this paper is in taking a first step toward increasing safety within an HRI scenario by integrating and evaluating a model that derives information about the human's engagement in an interaction based on a sequence of "joint action signals" humans naturally send when interacting with each other. Such signals include eye-head gaze orientation as a sign of a human's focus of attention and engagement in a task at hand.

Compared to a basic setting that estimates the state of the interaction between human and robot purely by using values related to the physical side of the handover (e.g. motor current and torque values), monitoring and considering the user's behaviour (which constitutes a "Theory of Mind" from the robot's perspective), as we demonstrated in our extended model, clearly increases the success rate of the handover. Thus, including human intention into the robot's decision process makes a robot-to-human object handover more robust and therefore safer. Note that we are fully aware of the fact that human intention on the basis of eye-head gaze orientation cannot be estimated with absolute certainty as humans might use gaze to deceive. However, as a first approximation eye-head gaze direction measures appear to be sufficiently reliable.

Future research directions include more detailed investigations into the cognitive aspects of joint attention in HRI. An important area for future work is the exploration of the "Theory of Mind" concept in the context of HRI, especially as a basis for developing more advanced social skills in personal robots. We have also started to develop techniques towards the verification and validation of HRI systems.

ACKNOWLEDGMENT

The authors wish to thank Alex Lenz and Sergey Skachek for their help with BERT2 and their insightful suggestions.

REFERENCES

[1] C. Vesper, S. Butterfill, G. Knoblich, and N. Sebanz, "A minimal architecture for joint action," *Neural Netw*, vol. 23, no. 8–9, pp. 998–1003, 2010.

[2] E. Birmingham and A. Kingstone, "Human social attention. a new look at past, present, and future investigations," *Annals of the New York Academy of Sciences*, vol. 1156, pp. 118–140, 2009.

[3] B. D. Corneil, *The Oxford Handbook of Eye Movements*. Oxford University Press, 2011, ch. Eye-head gaze shifts, pp. 303–322.

[4] "Principles of Robotics," http://www.epsrc.ac.uk, Sep. 2010.

[5] R. A. Brooks, "A robust layered control system for a mobile robot," Massachusetts Institute of Technology, Cambridge, MA, USA, Tech. Rep., 1985.

[6] E. Gat, R. P. Bonnasso, R. Murphy, and A. Press, "On Three-Layer Architectures," *Artificial Intelligence and Mobile Robotics*, pp. 195–210, 1997.

[7] G. Metta, L. Natale, S. Pathak, L. Pulina, and A. Tacchella, "Safe and effective learning: A case study," in *IEEE International Conference on Robotics and Automation*. IEEE, May 2010, pp. 4809–4814.

[8] S. Bensalem, M. Gallien, F. Ingrand, I. Kahloul, and T.-H. Nguyen, "Toward a More Dependable Software Architecture for Autonomous Robots," *IEEE Robotics and Automation Magazine*, vol. 16, no. 1, pp. 67–77, Mar. 2009.

[9] A. Basu, M. Gallien, C. Lesire, T. H. Nguyen, S. Bensalem, F. Ingrand, and J. Sifakis, "Incremental Component-Based Construction and Verification of a Robotic System," in *18th European Conference on Artificial Intelligence*, 2008, pp. 631–635.

[10] R. Muradore, D. Bresolin, L. Geretti, P. Fiorini, and T. Villa, "Robotic surgery," *Robotics Automation Magazine, IEEE*, vol. 18, no. 3, pp. 24–32, September 2011.

[11] D. Bresolin, L. D. Guglielmo, L. Geretti, and T. Villa, "Correct-by-construction code generation from hybrid automata specification," in *IWCMC*. IEEE, 2011, pp. 1660–1665.

[12] E. C. Grigore, K. Eder, A. Lenz, S. Skachek, A. G. Pipe, and C. Melhuish, "Towards safe human-robot interaction," in *12th Conference Towards Autonomous Robotic Systems*, vol. LNCS 6856. Springer, Sep. 2011, pp. 323–335.

[13] D. P. Carey, D. S. S., and M. Ietswaart, "Neuropsychological perspectives on eye-hand coordination in visually-guided reaching." *Prog Brain Res*, vol. 140, pp. 311–27, 2002.

[14] P. F. Ferrari, L. Bonini, and L. Fogassi, "From monkey mirror neurons to primate behaviours: Possible 'direct' and 'indirect' pathways," *Philos Trans R Soc Lond B Biol Sci*, vol. 364, no. 1528, pp. 2311–23, 2009.

[15] R. C. Schmidt, P. Fitzpatrick, R. Caron, and J. Mergeche, "Understanding social motor coordination," *Hum Mov Sci*, vol. 30, no. 5, pp. 834–45, Oct. 2011.

[16] W. P. Chan, C. A. Parker, H. M. Van der Loos, and E. A. Croft, "Grip forces and load forces in handovers: implications for designing human-robot handover controllers," in *7th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12. New York, USA: ACM, 2012, pp. 9–16.

[17] S. Shibata, B. Sahbi, K. Tanaka, and A. Shimizu, "An analysis of the process of handing over an object and its application to robot motions," in *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation*, vol. 1, Oct. 1997, pp. 64–69.

[18] M. Huber, A. Knoll, T. Brandt, and S. Glasauer, "Handing over a cube: spatial features of physical joint-action." in *17th IEEE International Symposium on Robot and Human Interactive Communication*, May 2008, pp. 107–112.

[19] S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt, "Interacting in time and space: Investigating human-human and human-robot joint action," in *19th IEEE International Symposium on Robot and Human Interactive Communication*, Sep. 2010, pp. 252–257.

[20] C. Becchio, L. Sartori, and U. Castiello, "Toward You: The Social Side of Actions," *Current Directions in Psychological Science*, vol. 19, no. 3, pp. 183–188, Jun. 2010.

[21] M. Cakmak, S. S. Srinivasa, M. K. Lee, S. Kiesler, and J. Forlizzi, "Using spatial and temporal contrast for fluent robot-human handovers," in *Proceedings of the 6th International Conference on Human-Robot Interaction*. New York, USA: ACM Press, 2011, p. 489.

[22] A. Edsinger and C. C. Kemp, "Human-Robot Interaction for Cooperative Manipulation: Handing Objects to One Another," in *16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2007, pp. 1167–1172.

[23] J. E. Hoffman and B. Subramaniam, "The role of visual attention in saccadic eye movements," *Perception and Psychophysics*, vol. 57, no. 6, pp. 787–795, 1995.

[24] S. Baron-Cohen, "The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology," in *Joint Attention: Its Origins and Role in Development*, C. C. Moore and P. J. Dunham, Eds. Lawrence Erlbaum, 1995, pp. 41–59.

[25] A. Frischen, A. P. Bayliss, and S. P. Tipper, "Gaze cueing of attention: Visual attention, social cognition, and individual differences," *Psychological Bulletin*, vol. 133, no. 4, pp. 694–724, Jul. 2007.

[26] A. Lenz, S. Skachek, K. Hamann, J. Steinwender, A. G. Pipe, and C. Melhuish, "The BERT2 infrastructure: An integrated system for the study of human-robot interaction," in *10th IEEE-RAS International Conference on Humanoid Robots*. IEEE, Dec. 2010, pp. 346–351.

[27] P. Fitzpatrick, G. Metta, and L. Natale, "Towards long-lived robot genes," *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 29–45, Jan. 2008.

[28] S. Sutton, R. Cole, J. D. Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki, P. Hosom, A. Kain, Johan, J. Wouters, D. Massaro, and M. Cohen, "Universal Speech Tools: The Cslu Toolkit," in *International Conference on Spoken Language Processing (ICSLP)*, 1998, pp. 3221–3224.

[29] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[30] C. X. Zhai, "A Brief Note on the Hidden Markov Models (HMM)," University of Illinois at Urbana-Champaign, Department of Computer Science, Tech. Rep., 2003.

[31] F. Paas, A. Renkl, and J. Sweller, "Cognitive load theory and instructional design: Recent developments," *Educational Psychologist*, vol. 38, no. 1, pp. 1–4, 2003.