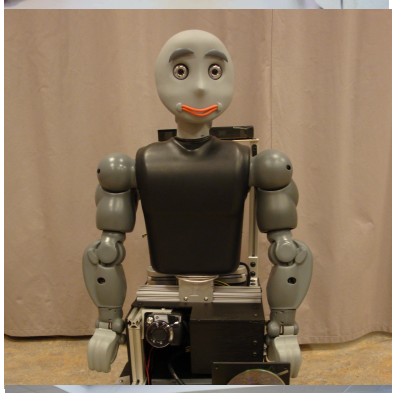
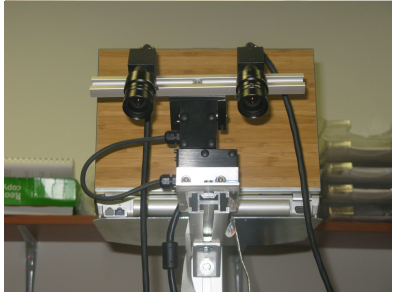
Proceedings of the ICRA 2010 Workshop on

Interactive Communication for Autonomous Intelligent Robots (ICAIR)

Making robots articulate what they understand, intend, and do.





Marc Hanheide
Hendrik Zender
(Eds.)

Workshop Organizers:

Marc Hanheide
University of Birmingham
School of Computer Science, Robotics and Cognitive Architectures Group
Birmingham, UK

Hendrik Zender
German Research Center for Artificial Intelligence (DFKI GmbH)
Language Technology Lab
Saarbrücken, Germany

Workshop Program Committee:

Tony Belpaeme
(University of Plymouth)

Shuzhi Sam Ge
(National University of Singapore,
International Journal of Social Robotics)

Patric Jensfelt
(KTH Stockholm)

Geert-Jan M. Kruijff
(DFKI Saarbrücken)

Matthias Scheutz
(Indiana University)

Marc Schröder
(DFKI Saarbrücken)

Adriana Tapus
(ENSTA Paris)

Ingrid Zukermann
(Monash University)

THE UNIVERSITY
OF BIRMINGHAM

German Research
Center for Artificial
Intelligence GmbH

ICRA 2010 Workshop W30-FrF
May 8, 2010, Anchorage, AK, USA

http://www.dfki.de/cosy/www/events/icair-icra2010

supported by the
EU-funded project CogX
(ICT – 215181 – CogX)

Time	Presentation	Authors / Presenter
09:15-09:30	*Introduction*: The role of feedback, articulation, and verbalisation	Workshop organizers: Hendrik Zender, Marc Hanheide
09:30-09:45	*Getting to know each other*	
	Track 1: Ideas, Foundations and Enabling Technologies	
09:45-10:30	*Invited Talk:* From explicit to implicit communication: is alignment the solution?	Britta Wrede, *Bielefeld University*
10:30-11:00	*coffee break*	
11:00-11:25	Speaker detection for conversational robots using synchrony between audio and video	Athanasios Noulas, Gwenn Englebienne, Bas Terwijn, Ben Kröse
11:25-11:45	Using hesitation gestures for safe and ethical human-robot interaction	AJung Moon, Boyd Panton, H.F.M. Van der Loos, E.A. Croft
11:45-12:05	Monitoring and guiding user attention and intention in human-robot interaction	Aaron St. Clair, Ross Mead, Maja J. Mataric
12:05-12:25	Generating multi-modal robot behavior based on a virtual agent framework	Maha Salem, Stefan Kopp, Ipke Wachsmuth, Frank Joublin
12:30-14:00	*lunch break*	
	Track 2: Interactive Systems	
14:00-14:45	*Invited talk:* The Autonomous City Explorer: Experiences from a recent test trial in the city center of Munich	Kolja Kühnlenz, *Technical University Munich*
14:45-15:05	Robot, tell me what you know about...?: Expressing robot's knowledge through interaction	Raquel Ros Espinoza, Akin Sisbot, Severin Lemaignan, Amit Pandey, Rachid Alami
15:05-15:30	A basic cognitive system for interactive curious learning of visual concepts	Daniel Skocaj, M. Janicek, M. Kristan, G.-J. M. Kruijff, A. Leonardis, P. Lison, A. Vrecko, M. Zillich
15:30-16:00	*coffee break*	
16:00-16:20	Identifying and resolving ambiguities within joint movement scenarios in HRI	Maryamossadat N. Mahani, Elin Anna Topp
16:20-16:40	The Curious Robot learns grasping in multi-modal interaction	Ingo Lütkebohle, Julia Peltason, Robert Haschke, Britta Wrede, Sven Wachsmuth
16:40-17:30	*Plenary and Closing Remarks*	

Contents

Introduction

Making robots articulate what they understand, intend, and do.

Human-robot interaction is becoming more and more complex through the growing number of abilities, both cognitive and physical, available to today's robots and through their resulting flexibility. At the same time, lay persons should be able to interact with robots in order to pursue the vision of a robot in every home. Though a lot of progress is apparent in the different fields in robotics with regard to learning, autonomous behaviours, safe navigation, and manipulation, the interface with the human user is quite often rather neglected. Many studies have been conducted unveiling the importance of properly designed adaptive human-robot interaction strategies and appropriate feedback, in particular. With interaction becoming more complex it is equally becoming more important to move beyond command style interfaces and equip robots with abilities to actually express and verbalise what they are doing, what their current problems might be and how they see the world. These interactive abilities have been shown to facilitate more effective and efficient interaction with humans using mostly natural modalities, but also robot-specific ones, such as visualisation techniques.

Invited Talks

The Autonomous City Explorer: Experiences from a recent test trial in the city center of Munich

Kolja Kühnlenz (Technical University Munich)

Abstract Future personal robots in everyday real-world settings will have to face the challenge that there will always be knowledge gaps. A priori knowledge may not be available in all situations and learning requires trials, which also may not be feasible in any case. In order to overcome such drawbacks, we believe that a crucial capability of tomorrow's robot assistants will be to assess their knowledge towards gaps and to be able to fill those by interaction with humans. In this talk, recent results of the Autonomous City Explorer (ACE) project will be presented. In this project, an autonomous robot managed to find its 1,5km way from the main campus of TU Munich to the city center of Munich by asking pedestrians for directions. ACE was developed in the context of a pilot project exploring the feasibility of personal assistance robots in terms of human acceptance, which are capable of extending their knowledge not only by means of cognition but also by means of humanlike communication in real-world settings. To fill gaps in its directional knowledge, ACE is capable of actively approaching humans and initiating interaction situations, retrieving directions from human pointing gestures and converting this information into an algorithmic plan, which finally is executable in terms of conventional means of robot navigation.

About the speaker Kolja Kühnlenz is currently a Senior Lecturer at the Institute of Automatic Control Engineering (LSR) and Carl von Linde Junior Fellow at the Institute for Advanced Study, Technische Universität München, Munich, Germany. He is director of the Dynamic Vision Research Laboratory at LSR with currently 7 PhD students. His research interests include Robot Vision, Visual Servoing, High-Speed Vision, Attention, Bio-inspired Vision, Humanoid Robots, Human-Robot Interaction, Emotions, and Sociable Systems – with a strong focus on real-world applications of (social) robots.

From explicit to implicit communication: is alignment the solution?

Britta Wrede (Bielefeld University)

Abstract In recent years the theory of grounding according to which participants explicitely negotiate what they have understood and thus build a common ground has been challenged by the idea of a mechanistic view of understanding, Alignment. The latter idea is based on the observation that in task-oriented interactions communication partners tend to align their surface representations (e.g. lexical or syntactic choice) in an implicit way which apparently helps to align their underlying situation models and thus facilitates mutual understanding. In this talk, Britta Wrede will present some experimental analyses of human-robot interaction where misunderstandings occur that are often caused by implicit signals from the robot which are interpreted by the human in a communicative way. It will be discussed if such implicit mechanisms of understanding can be useful in human-robot interaction.

About the speaker Britta Wrede is head of the research group Hybrid Society within the Institute for Cognition and Robotics (CoR-Lab) at Bielefeld University. She received her Masters degree in Computational Linguistics and the Ph.D. degree (Dr.-Ing.) in computer science from Bielefeld University in 1999 and 2002, respectivley. From 2002 till 2003 she pursued a PostDoc program of the DAAD at the speech group of the International Computer Science Institute (ICSI) in Berkeley, USA. In 2003 she rejoined the Applied Informatics Group at Bielefeld University and was involved in several EU and national (DFG, BMBF) projects. Since 2008 she is heading her own research group at the CoR-Lab. Her research interests include speech recognition, prosodic and acoustic speech analysis for propositional and affective processing, and dialog modeling as well as human-robot interaction. Her current research focuses on the integration of multi-modal information as a basis to bootstrap speech and action learning in a tutoring scenario.

Speaker detection for conversational robots using synchrony between audio and video

Athanasios Noulas, Gwenn Englebienne, Bas Terwijn and Ben Kröse
Informatics Institute, University of Amsterdam,
The Netherlands
b.j.a.krose@uva.nl

***Abstract*— This paper compares different methods for detecting the speaking person when multiple persons are interacting with a robot. We evaluate the state-of-the-art speaker detection methods on the iCat robot. These methods use the synchrony between audio and video to locate the most probable speaker. We compare them to simple motion-based speaker detection and present a simple heuristic with low computational requirements, which performs equally well to the audiovisual methods in a set of multiperson recordings with a fraction of the computational cost, thus making real-time interaction possible.**

I. Introduction

Embodied conversational agents, whether physical robots or screen agents will play an important role in future man-machine interfacing. Such systems must intelligently interpret the voices they perceive, even in a multi speaker setting. Furthermore, the embodied agent must react in a 'social' way to the humans interacting with it. This means that conventions that play a role in man-to-man conversation must also be implemented in man-to-robot conversations. The roles of person detection, gaze control and eye contact have been extensively studied in man-robot interaction [18], [25] and the interaction with animated faces [10], [6]. However, most of this work focuses on the detection of, and interaction with, humans in a single user situation. In many of the foreseen applications such as robots in museum or exhibitions, robots in care-for-elderly (see figure 1), office robots and entertainment robots, these systems have to interact with multiple humans simultaneously.

As a part of this task we address the problem of detecting the person that is speaking in a situation where multiple persons interact with the robot. Speaker detection in such applications is done along two lines. One line is the use of multiple microphones to detect the location of the speaker [1],[13]. The second line of research focuses on combining the audio based localization with detection of the location of humans by other means such a vision or laser [16].

In this paper we explore how speaker detection methods, developed for the automatic analysis of multimodal information streams such as meeting videos of news broadcasts, can be applied to a robotics application. These methods use *synchrony* between audio and video to find the speaking person. We compare these methods with a simple, very fast, ad-hoc method that is based on motion detection only.

This work was partly supported by MultimediaN

Fig. 1. An example of elderly care where multiple persons interact with a robot.

We will first present related work in the field of speaker detection and then briefly present the framework that we developed for speaker diarization in multimodal streams. Section VI presents the experimental results using the i-Cat robot while sections VIII and IX discuss the results and present the conclusions of this work.

II. Related work

In robotics research, speaker detection is usually considered as a sound source localisation and tracking problem in which audio information from two or more microphones is used. The aim of speaker localisation is mostly the enhancement of the speech signal [1]. This can be done by for example adaptive beamforming as described by Beh *et al.* [3]. Nakadai *et al.* [19] present the active direction-pass filter to separate sounds originating from the specified direction with a pair of microphones. Not only do they use a microphone array to localise audio, but they also vision cues in the form of face detection and stereo vision. The results show that the vision cues are more accurate than audio cues for localising the speaker. Note that, in this case, there is only a single speaker.

In multispeaker situations the data association problem has to be solved. Klaassen *et al.* [13] use a joint probabilistic

data association filter to localise two speakers using audio from two microphones. The detected formants are voice specific features, while Generalised Cross-Correlation derives the position-specific features from two microphone signals. Results show that although the localisation from audio is extremely inaccurate, the voice features allow for effective tracking. In [16], the audio based localisation is combined with a localisation of the legs of the speakers from a laser range finder and the detected head. A set of heuristic rules was designed to identify the speaker.

Most of the work done in audiovisual speaker detection is not done in robotics but in application fields like video conferencing or improving human computer interaction (HCI), in which the approaches focus on using the *synchrony* between the audio and video stream. Solutions come in two categories, namely the approaches based on Mutual Information (MI) [9], [20] and the approaches based on a matching algorithm [12], [2].

The approaches based on MI extract low-level features such as pixel intensities from the video and energy from the audio. Then, they implicitly assume that the MI between the audio and video features reflects audiovisual synchrony: the higher the MI, the more synchronised the original streams are. The derived MI measurements are used to select the image region that contains the speaker.

The approaches based on a matching algorithm process the audio and video signals extensively in order to extract low-dimensional high-level features such as the detection of sudden changes in the audio stream, or the acceleration of distinctive visual features. The synchrony between audio and video is calculated with some ad hoc distance measure, on the basis of which the speaker is selected.

The MI-based approaches are considered more suitable for speaker detection, since they are robust to environmental noises and do not require any elaborate feature extraction. Furthermore, MI-based approaches have been evaluated in multiple subject experiments in [8] and extensive monologue and speaker detection experiments on publicly available data sets in [11]. The matching algorithm-based approaches has only been applied in two speaker scenarios and only reported qualitative results[2], [12].

However, the matching algorithm results are particularly interesting for robotics: Barzelay et al. uses the detected feature to perform to source separation [2], and, in the current context, this can be applied to clear the voice of the speaker from background noise.

Both the MI-based and the matching algorithm-based approaches have been applied on high-quality audiovisual recordings and static cameras. These recordings were processed off-line, and all algorithms involve time consuming computations. The contrast is therefore sharp with typical robotics applications, where low quality recordings, moving cameras and real-time requirements are the norm.

In this work we test different speaker detection methods on the i-Cat robot. We implement the MI-based, and matching algorithm-based speaker detection as it was introduced in [9] and [2] respectively. We evaluate the results of these methods in recordings made through the i-Cat's camera and microphones, containing two to four speakers with a moving or static robot, and we compare these results to a simple motion detection-based method.

III. Motion-based Speaker Detection

Speaker detection based on motion assumes that the speaker will move in order to speak. This assumption is inspired not only by the facial actuators required for normal speech generation, but also because speakers naturally tend to rely on non-verbal communication methods in conjunction with verbal communication [5], [15], *e.g.*, nodding, conversational hand gestures, facial expressions, *etc*.

In a very simple yet surprisingly powerful approach, we rely on the difference between consecutive frames to detect the speaker. The algorithm consists of the following steps, which are executed for each consecutive pair of frames: (1) *Face detection:* detect the faces in the current frame, using a standard algorithm such as the Viola-Jones face detector [24]; (2) *Difference:* subtract the previous frame from the current frame within the resulting face regions; (2b) *Thresholding:* count the number of pixels for which the difference is above the threshold. For our experiments, we have chosen to set the value at $1/5$ of the dynamical range of the pixel intensities. Finally (3) *Selection:* select the face area which contains the largest total difference as the speaker.

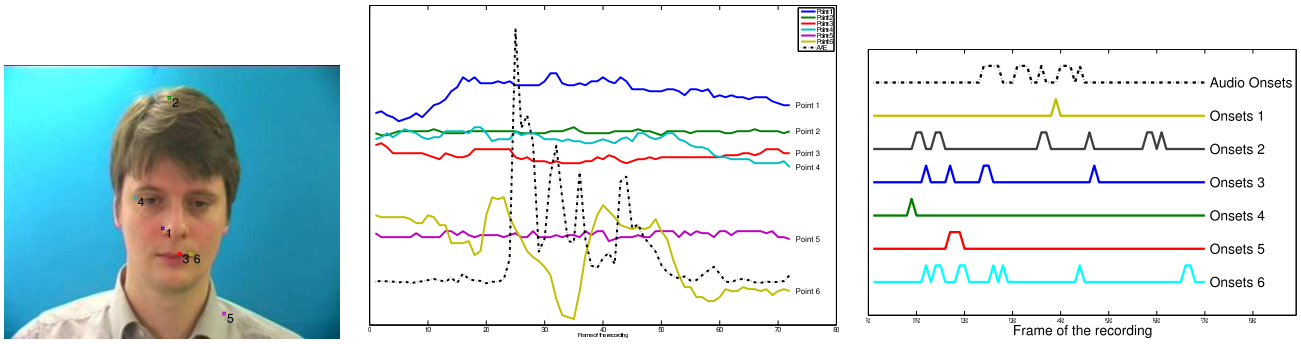
The main advantage of this algorithm is that it can be implemented very efficiently using the MIMD instructions present in contemporary processors, and can hence run much faster than in real time for the image resolution used in our experiments. The worst-case performance occurs in the hypothetical case where the complete frame is selected as a face. For the image resolution used in our experiment (*viz.* 320×240 intensity pixels) the difference operation then requires $15\,\mu$s, while the thresholding requires $12\,\mu$s on a 2.5GHz Core2 processor. On the much slower, low power 900MHz celeron processor of a netbook, these operations are still performed in less than $80\,\mu$s and 45μs, respectively. At these speeds, the speaker detection is essentially for free.

IV. Synchrony Based Speaker Detection

Speaker detection based on synchrony assumes that the person appearing most synchronised to the audio stream is the speaker. In practice, this is performed in three steps. In the first step, a *face detection* algorithm detects the faces in the frame. In the second step, the face regions are evaluated using the *synchrony detection* methods, which return either a measure of synchrony or the location of the visual feature appearing most synchronised to the audio stream. Based on the output of the synchrony detection method, a face is *selected* as the speaker. Synchrony can be detected using a method based on MI or a matching algorithm.

A. MI-based Methods

MI was first proposed for synchrony detection in the work of Hershey and Movellan [9], where it is assumed that MI between the audio and video features reflects synchrony



(a) Example frame (b) Pixel values and AAE (c) Video and audio onsets

Fig. 2. On the left, an example frame from a video sequence with 6 pixels selected, coming from the nose, the hair, the eye, the shirt and the lips of the person. On the middle the grey-scale value variation for the selected pixels over 72 frames, as well as the average acoustic energy of the audio stream over the same period (dashed line). On the right the onsets for the features corresponding for the pixels depicted in 2(a), and the onsets for the audio.

between the audio and video modalities. Intuitively, MI between variables $\mathbf{X}$ and $\mathbf{Y}$ measures the information about $\mathbf{X}$ that is provided by $\mathbf{Y}$. It is denoted as $MI(\mathbf{X};\mathbf{Y})$ and it is given by:

$$MI(\mathbf{X};\mathbf{Y}) = \int_{\mathbf{X}}\int_{\mathbf{Y}} p(\mathbf{x},\mathbf{y}) \log\left(\frac{p(\mathbf{x},\mathbf{y})}{p(\mathbf{x})\,p(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \quad (1)$$

Hershey and Movellan suggest the estimation of the MI between the pixel values and the average acoustic energy of the audio stream. In general, MI can not be computed explicitly in closed form. However, assuming that variables $\mathbf{X}$ and $\mathbf{Y}$ are Normally distributed, there exists a closed-form solution of their MI:

$$MI(\mathbf{X};\mathbf{Y}) = \frac{1}{2}\log\left(\frac{|\Sigma_{\mathbf{X}}||\Sigma_{\mathbf{Y}}|}{|\Sigma_{\mathbf{XY}}|}\right) \quad (2)$$

where $\Sigma_{\mathbf{X}}$ and $\Sigma_{\mathbf{Y}}$ are the covariances of the distributions of the variables $\mathbf{X}$ and $\mathbf{Y}$ respectively and $\Sigma_{\mathbf{XY}}$ the covariance of their joint distribution.

In our experiments, the MI was estimated between the intensity variation of each pixel in the face regions and the Average Acoustic Energy (AAE) of the corresponding audio stream. We use seven frames to compute the MI, which corresponds to 0.7 seconds of data, a choice similar to that of the original paper [9]. The AAE of an audio window is estimated as the sum of the absolute values of its samples. The size of the audio window is equal to the frame size, i.e. 100 ms.

In order to acquire a measure for the face window in the frame, the average MI of the pixels of that area is used. In figure 2(a) an example frame of a speaking person is presented. To compare with the matching algorithm we manually selected 6 different pixels from the face region that are also salient features in the matching algorithm. The gray-scale values of these pixels as well as the AAE of the corresponding audio stream are plotted in figure 2(b). The pixel coming from the edge of the lips (point 6) exhibits the highest variation while the rest of the pixels exhibit little variation. Notice that the corresponding audio stream also exhibits variation at the same time that the pixel coming from the edge of the mouth does. However, a nearby pixel (point 3) does not exhibit a similar behaviour, because there is little image texture around that pixel.

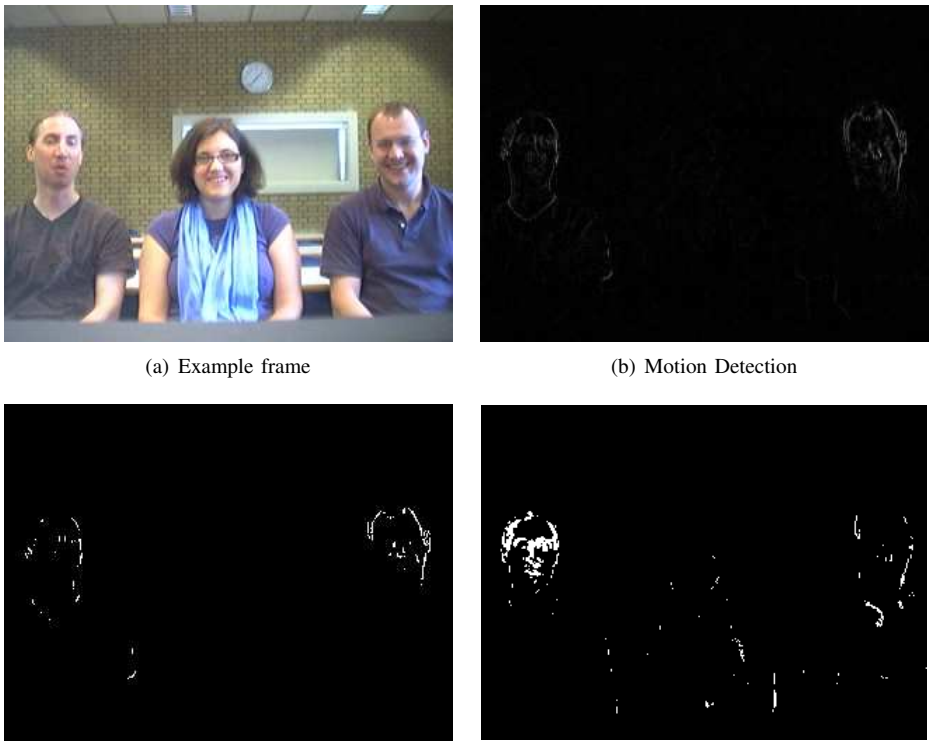
Given one or more windows from the face detector, we compute the average MI between each window's pixels and the AAE. The face window which produces the highest MI measurement is expected to be the most synchronised to the audio stream and it is selected as the speaker.

B. Matching algorithm -based methods

Previous research also explored synchrony detection on high-level features, *i.e.*, features for the extraction of which extensive processing of the input signals is required. In this line, Barzelay and Schechner in [2], which extends the earlier work of Kidron *et al.* [12], seek correspondence between *significant* features in the audio and video streams. This is a choice motivated by biological neural systems research concluding that cross-modality association is based on salient features [7]. In synchrony detection, the characteristics of significant features are saliency, reliable detection and high correlation in the audio and video modality. In the work of Barzelay and Schechner, the features regarded significant are *onsets* in the *video* and *audio* modality. Onsets in the video and audio modality are points in the stream where each signal exhibits strong temporal variation [2].

In the *video* modality, the first step is to detect features that can be tracked over multiple frames. In the works mentioned above, Kanade-Lucas-Tomasi (KLT) features are used. KLT features are located by examining the minimum eigenvalue of each two by two gradient matrix, and they are tracked using a Newton-Raphson method of minimising the difference between two consecutive windows. Multi-resolution tracking allows for relatively large displacements between images. The original idea for such tracking chosen dates back to 1981 and the work of Lucas and Kanade [17], and the implementation used in our experiments was further developed in the works of Tomasi and Kanade [22] and Shi and Tomasi [21]. In figure 2(c) the onsets for the features corresponding to the points of figure 2(a) are plotted.

In order to decide when an onset occurs, each feature i is tracked independently. The magnitude of the feature's



(a) Example frame (b) Motion Detection

(c) Motion Detection with threshold (d) Mutual Information Image

Fig. 3. On the left, an example frame from a video sequence from the iCat camera. From left to right we visualise motion detection, motion detection with threshold and Mutual Information. Brighter values in the visualisation correspond to higher values. Note that in this case the person on the left is speaking, something only detectable in the MI image.

acceleration at frame t is measured, thresholded and temporally pruned. This results in a binary vector $\mathbf{v}_i$ for each feature i,, where element $v_i(t)$ is one if feature i has high acceleration at t and zero otherwise. In figure IV the onset vectors of six selected features are shown — the selected features correspond to the points whose gray scale value variation was shown in figure 2(b).

In the audio modality, onset detection is a well-studied problem, see for example the tutorial of Bello *et al.* [4]. In our experiments the detected onsets were based on psychoacoustic knowledge as described in the work of Klapuri [14]. In short, the initial audio signal is divided into 21 non overlapping frequency bands. Onset detection is performed in each band independently, by locating the peaks in the first derivative of the logarithm of the amplitude envelope. In the final step, detected onsets in all banks are gathered, and the sum of the onset intensities is estimated. In parallel to the processing of the video modality, the total intensity for each candidate onset is thresholded to provide the onset locations in a vector $\mathbf{a}(t)$. The detected onsets for the AAE plotted in figure 2(b) are shown in figure 2(c).

The matching algorithm performs synchrony detection in the onset space. The matching criterion is defined as:

$$L(i) = 2\left[\mathbf{a}^T \mathbf{v}_i\right] - \mathbf{1}\mathbf{v}_i \tag{3}$$

where $\mathbf{1}$ is a column vector with all elements equal to one.

The feature point with the highest value in the matching criterion is selected as the source of the corresponding audio stream. Given one or more windows from the face detector, the feature in these windows with the highest matching criterion is selected as the speaker.Note that the matching algorithm depends to a large extend on the quality of the audiovisual material.

V. Tests and implementation on a conversational robot

The proposed algorithms for on-line speaker detection have been tested using videos taken from the i-Cat robot. The iCat robot is a robotic character developed by Philips Research for HCI research [23]. It is equipped with a camera mounted on its head that can pan and tilt. The camera is a simple webcam with a 320x240 resolution. For the audio recording a single Buddy DesktopMic^{tm} mono microphone was used. The iCat is controlled using the Open Platform for Personal Robotics (OPPR) version 2.0.5. From version 2.0 onward, the so called behaviors allow for relative head movement fast enough for tracking. The behaviors are implemented by LUA script and give direct access to the iCats servos. Figure 3 presents an example frame from an iCat video 3(a), the corresponding motion detection 3(b), motion detection with threshold 3(c) and the corresponding Mutual Information Image 3(d).

We also implemented the motion based speaker detection and the MI based speaker detection on the iCat. The iCat is

controlled by two laptops[1]. The first laptop has two 2.33Ghz processors and 2GB of RAM, and analyses of the audio-visual signal processing. The second laptop has one 3Ghz processor and 512MB of RAM, and it is responsible for controlling the iCat. The iCat records records video at 10fps. Face detection using the Viola Jones face detector can be performed on the used processors with a speed of 25fps, which is much faster than the video rate of iCat. The motion detection, which has negligible cost, will not influence the video processing. Methods based on MI are more time consuming. Under the current implementation, and estimating the MI only for the face regions, we achieve frame rates of approximately 15fps. Considering that the iCat camera records records video at 10fps this is practically real time, but no generalisation holds for high quality recordings. Finally, the Matching Algorithm approach can not be performed in real time on these processors, and it is evaluated off-line on the recorded video.

VI. Experiments

Experiments are conducted on videos where multiple persons facing the iCat robot are speaking in turn. Videos were recorded with 2, 3 and 4 persons. For each number of persons a video was recorded with both the iCat being static and dynamic. In dynamic mode the iCats head was moving left and right as to gaze at persons in front of it. This is done to simulate the behavior of a social robot looking at the persons in front of it complicating the speaker detection problem. In this mode, persons at the sides leave and reenter the field of view as the camera moves. In static mode all persons remain in the field of view at all time. The movements consists of panning the head between the left, center and right position moving about 25 degrees each step. A step was performed after a random interval of 2 to 9 seconds. The movement itself took about 0.5 seconds and causes the frames to be blurred during that interval. This means that the performance degrades because the face detection is more difficult.

During recording the persons were seated in a row at a distance of about 3 meters from the iCat and in turn recited from a theater play script in which a turn would last between 1 and 40 seconds. The same script was used in all videos. The shortest video lasts 409 seconds and the longest 496 seconds. All 6 videos were recorded at the same location in good lighting and noise conditions with a framerate of 10 frames per second and a high audio bitrate.

The ground truth for the experiments was established by manually going through the videos and annotating who was speaking at each frame. Hereby the movement of lips was the most important cue. We had to rely on only audio data when the speaking person was outside of the field of view. Some ambiguity is present when a person pauses shortly while reciting. Non-speaking sounds such as laughing and coughing are considered speaking when they originate from the speaking persons during or shortly before and after that person is speaking.

[1]Communication between tracking software and a behavior was done via sockets as we found the Philips Dynamic Module Library (DML) introduced additional latencies.

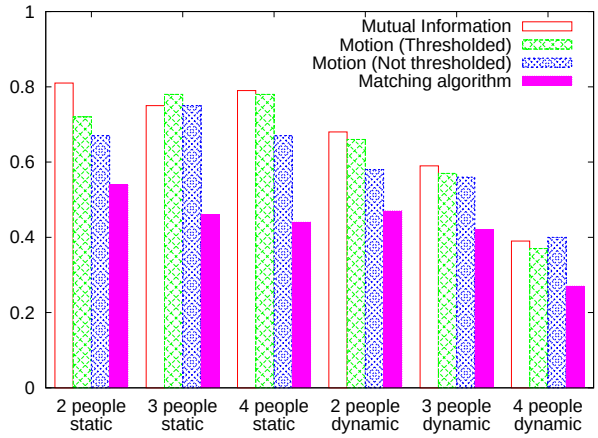


Fig. 4. Bar diagram of the accuracy for different methods

The accuracy of a method corresponds to the number of frames in which the speaker is correctly detected. We measured this by measuring the number of false detections and comparing this to the total number of frames where someone is speaking:

$$\text{Accuracy} = 1 - \frac{\text{Error Frames}}{\text{Total Frames}} \tag{4}$$

note that the frames containing silence are not considered in the error rate measurement.

VII. Results

Figure 4 plots the accuracies of different methods on different videos. The performance of the MI method is slightly better than that of motion detection in most of the cases. This difference is not statistically significant, it is however systematic in the videos of the dynamic camera, which are common in the robotics applications. The matching algorithm method, which produces state-of-the-art results in high-quality recordings that are processed off-line, performs significantly worse in these experiments. Note that for the static camera the face detection was perfect, yielding a best performance of 100%.

In the experiments performed with a dynamic camera, the speaker is often not visible, and furthermore face detection is low during the motion of the camera. Consequently, the best performance does not correspond to 100% accuracy. The optimal possible performance, which corresponds to an accuracy related to the percentage of frames where the speaker is visible, is 78% for the 2-people, 68% for the 3-people and 57% for the 4-people situation.

VIII. Discussion

In static camera videos, speaker detection based on MI performs better than the other methods. When a threshold is used in motion detection the results improve systematically, because the noise coming from the recording equipment is successfully filtered out. This is very insightful: Mutual

Information is higher for pixels with high variance when someone is speaking, *i.e.* the same pixels that will have high differences and will be therefore kept during thresholding.

In a dynamic camera, speaker detection based on MI achieves the best overall results. The consequences of thresholding are more systematic in this case. When it is beneficial to threshold the pixel values, MI and motion detection with threshold perform better than motion detection without threshold. When it is not beneficial, they perform worse. Moreover, Mutual Information performs better because it applies a complex thresholding that takes into account the variation in the audio modality, rather than looking at the video modality alone.

Finally, speaker detection based on the matching algorithm produces the worse results. This is due to the low quality of the recording in terms of video analysis, frame rate and audio quality. The method proposed by Barzelay et al. is not directly transferable to a conversational robot [2].

IX. Conclusions

We showed that visual information can successfully be used for speaker detection by a robot in a multiple speaker setting. Our results show that there is no need for two microphones or a microphone array for speaker localization, as long as the speakers are visible.

Our comparison between the simple method based on motion and the more involved methods based on audiovisual synchrony, showed that these latter methods did not perform significantly better. Using a single microphone improves marginally over the single video modality approach. Furthermore, the matching algorithm performed substantially worse than the motion or MI method.

Apparently these results are different from findings in multimodal speaker diarization. A difference between the two application areas is that in robotics the quality of the audiovisual recordings is generally low. For a fair comparison it is needed to test all three methods on high quality audiovisual material such as the AMI meeting corpus. However, for real time operation we have to adhere to fast methods such as the motion based or MI based methods.

Future work will focus on user studies with the methods. We are able to run the motion based speaker detection and the MI based speaker detection on the iCat and acceptance studies will be carried out. In this way we will be able to study the actual 'conversational' skills of the system instead of just speaker detection.

References

[1] F. Asano, M. Goto, K. Itou, and H. Asoh. Real-time sound source localization and separation system and its application to automatic speech recognition. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[2] Z. Barzelay and YY Schechner. Harmony in motion. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.

[3] Jounghoon Beh, Taekjin Lee, Inho Lee, Hyunsoo Kim, Sungjoo Ahn, and Hanseok Ko. Combining acoustic echo cancellation and adaptive beamforming for achieving robust speech interface in mobile robot. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1693–1698, 2008.

[4] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davis, and Mark B. Sandler. A tutorial on onset detection in musical signals. *IEEE Transactions on speech and signal processing*, 13(5):1035–1047, 2005.

[5] Justine Cassell and Kristinn R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence: An International Journal*, 13(4):519, 1999.

[6] A. Colburn, M.F. Cohen, and S. Drucker. The role of eye gaze in avatar mediated conversational interfaces. *Microsoft Research Report*, 81:2000, 2000.

[7] Waka Fujisaki and Shin'ya Nishida. Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research*, 166(3-4):455–464, October 2005.

[8] Iyengar Giridharan, Nock Harriet J., and Neti Chalapathy. Audio-visual synchrony for detection of monologues in video archives. In *International Conference on Multimedia and Expo*, pages 329–332. IEEE Computer Society, 2003.

[9] J. Hershey and J. Movellan. Audio-vision: Using audio-visual synchrony to locate sounds. *Advances in Neural Information Processing Systems*, 12:813–819, 2000.

[10] D. Heylen, I. van Es, B. van Dijk, and A. Nijholt. Experimenting with the gaze of a conversational agent. *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*, page 93, 2002.

[11] Nock Harriet J., Iyengar Giridharan, and Neti Chalapathy. Multimodal processing by finding common cause. *Communications of the ACM*, 47(1):51–56, 2004.

[12] E. Kidron, YY Schechner, and M. Elad. Pixels that sound. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 1, 2005.

[13] G. Klaassen, W. Zajdel, and B.J.A. Kröse. Speech-based localization of multiple persons for an interface robot. In *Proc. of IEEE Int. Conference on Computational Intelligence in Robotics and Automation (CIRA2005)*, pages 47–52, 2005.

[14] Anssi Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3089–3092. IEEE Computer Society, 1999.

[15] Robert M. Krauss, Robert A. Dushay, Yishiu Chen, and Frances Rauscher. The communicative value of conversational hand gesture. *Journal of Experimental Social Psychology*, 31(6):533–552, November 1995.

[16] S. Lang, M. Kleinehagenbrock, J. Fritsch, G. A. Fink, and G. Sagerer. Detection of communication partners from a mobile robot. In *in Proc. of the 4th Workshop on Dynamic Perception*, pages 183–188, 2002.

[17] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with application to stereo vision. In *Internation Conference on Artificial Intelligence*, pages 121–130, 1981.

[18] D. Miyauchi, A. Nakamura, and Y. Kuno. Bidirectional eye contact for human-robot communication. *IEICE TRANSACTIONS on Information and Systems*, (11):2509–2516, 2005.

[19] Kazuhiro Nakadai, Hiroshi G. Okuno, Hiroaki Kitano, Hiroshi G. Okuno, and Hiroaki Kitano. Real-time sound source localization and separation for robot audition. In *in Proceedings IEEE International Conference on Spoken Language Processing, 2002*, pages 193–196, 2002.

[20] H.J. Nock, G. Iyengar, and C. Neti. Speaker localisation using audio-visual synchrony: An empirical study. *Lecture Notes in Computer Science*, pages 488–499, 2003.

[21] Jianhao Shi and Carlo Tomasi. Good features to track. In *Conference on Computer Vision and Pattern Recognition*, pages 593–600. IEEE Computer Society, 1994.

[22] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University, April 1991.

[23] Albert van Breemen, Xue Yan, and Bernt Meerbeek. icat: an animated user-interface robot with personality. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, pages 143–144. ACM, 2005.

[24] Paul Viola and Michael Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[25] Y. Yoshikawa, K. Shinozawa, H. Ishiguro, N. Hagita, and T. Miyamoto. Responsive robot gaze to interaction partner. In *Proceedings of robotics: Science and systems*, 2006.

Using Hesitation Gestures for Safe and Ethical Human-Robot Interaction

AJung Moon, Boyd Panton, H.F.M. Van der Loos and E.A. Croft

Abstract— **Safe interaction with non-expert users is increasingly important in the development of robotic assistants. Ethical "codes" can serve as a guide as to how this interaction should take place with lay users in non-structured environments. Such codes suggest that robots should behave in a way that is intuitive to users. Previous research has demonstrated that the implicit channel is useful for intuitive human-robot interaction. Our work described in this position paper investigates how a robot should behave when it is uncertain of its human partner's intentions. In this context, uncertainties arising in human-robot shared-tasks should be made transparent to a human user. We posit that hesitant hand motion used by people and animals is a natural modality for a robot to communicate uncertainty. To test our hypothesis we propose to characterize and implement human hesitation gestures onto a robot, and investigate its ability to communicate uncertainty.**

I. INTRODUCTION

EVERY year, increasingly sophisticated robots intended for personal and service applications are showcased. A society in which people routinely interact with robots in home and office environments, while sharing working space, tasks and objects, is becoming a realizable and anticipated future. As a result, increases in autonomy, ability, and complexity of robots are inevitable and gradually requiring more autonomous decision-making capability with minimal human intervention. This raises concerns regarding an expected "code" of conduct that guides robot behavior, namely, robot ethics. In this position paper we uphold the argument of others [1] that ethical robot behavior, as it pertains to interactions with humans, must be considered in order to successfully integrate domestic robots into our society. Unlike traditional ethical questions, which are constantly under debate, robot behavior ethics within a given context can be framed in terms of human safety and social norm adherence [2].

Robots for service and domestic applications pose interesting challenges to issues of safety and ethics [3]. These robots frequently encounter new, uncertain and conflicting situations where any resulting indecision or inaction can bring negative consequences to the user. In such cases, it is important for a robot to clearly communicate its intentions to the user. Take, for example, the annoyance a user may experience with a wheelchair robot when attempting to hang a picture on a wall. Detection of an imminent collision with the wall coupled with the user's command to move forward may introduce uncertainty to the wheelchair's controller. Subsequently, the user may be unable to achieve the desired goal due to inaction or indecision by the robot, and the inability to read the robot's internal state only adds to the frustration experienced by the user. With the possibility that unresolved uncertainties can result in dire consequences, Van der Loos [4] advocates that increase in complexity of robots should be followed by increase in transparency of robot intention in order for human-robot interaction (HRI) to be safe and ethical.

We posit that the appropriate action of a robot, when faced with uncertainty in an interaction, is to unambiguously demonstrate its internal state. Thus, we hypothesize that such transparency of the robot's inner state can improve user perception of robots. We also postulate that such interaction can initiate a human-friendly human-robot mediation process where the two agents can collaboratively solve the conflict and clarify the uncertainty.

Inspired by the body of work on implicit interaction [5]-[9], which collectively validates the use of nonverbal gesture as an effective communication and interaction mechanism in HRI, we are interested in studying whether a robot's state of uncertainty can be communicated to users via nonverbal gestures. In our study we take the exemplar case of two people noticing that they are reaching for the same object simultaneously. Our pilot studies have shown occurrences of sudden halts or jerky motions of participants' hands before one person yields or persists to resolve the uncertainty regarding who gets the object. Ultimately, with the proposed approach outlined in this paper, the outcome of our study will increase the understanding of how nonverbal gestures such as hesitations can be effective and appropriate in HRI.

II. BACKGROUND

A. Hesitation and Uncertainty

Existing work in psychology indicates that cognitive or internal state of uncertainties and conflicts in animals and humans are often expressed in terms of nonverbal gestures. Such nonverbal behaviors include shrugs, frowns, palm-up gestures and self-touch gestures [10]. Some causes of hesitant nonverbal behaviors are confusion [8], cognitive conflicts [11], difficulty in cognitive processing [12] and reluctance to act [13]. These sources of hesitation manifest themselves in multiple forms of resultant gestures. The previously described jerky motion between two people reaching for the same object arises from cognitive conflict

This work was supported by the Natural Sciences and Engineering Research Council of Canada.

A. Moon ajung@amoon.ca
B. Panton bepanton@engmail.uwaterloo.ca
H.F.M. Van der Loos vdl@mech.ubc.ca
E.A. Croft ecroft@mech.ubc.ca
The authors are with the Department of Mechanical Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

and reluctance to act. We label this kinesthetic gesture a 'hesitation gesture', and it is shown in Figure 1. We are currently investigating hesitation as a potential nonverbal robot gesture that can convey the robot's state of uncertainty to its human collaborator in a human-robot shared-task (HRST) environment.

(a)

(b)

Fig. 1. Hesitation gesture in a human-human shared-task

B. Nonverbal Gestures in HRI

Nonverbal gestures as interaction mechanisms in HRI have been studied in various contexts, types of robots, and modalities [5], [14]-[16]. Among the most studied are gestures used to regulate the flow of conversation between robots and humans [17], [18], and human-robot proxemics [19], [20]. Several studies have investigated the connection between non-verbal gestures and a robot's internal states [15], [21], [22]. However, these studies focused on the expression of emotional state. Nonverbal gestures used to communicate or express the cognitive state of a robot to a person remains relatively unexplored.

As previously mentioned, there are numerous hesitation gestures involved in expressing uncertainty. Breazeal's work on nonverbal robot gestures focused on expression of uncertainties arising from confusing human commands [8]. This study involved a teamwork scenario in which the human took a supervisory rather than collaborative role, instructing the robot to take specific actions. The robot expressed its internal state of uncertainty using shrugs. Her work provides strong evidence that use of nonverbal gestures rather than voice to render a robot's internal state transparent can be effective and helpful in improving task performance with lay users.

However, uncertainty due to cognitive conflict rather than confusion about a command occurs when a robot is interacting with a human as a near-equal partner. In our exemplar case, uncertainty arises regarding who should yield, and how the cognitive conflict between the desire to get the object and the need to meet social norms in being polite to another person is resolved. The gesture manifested from this type of uncertainty is the focus of our study, which we believe will have an impact in creating a human-friendly HRST for lay users when the robot is a near-equal partner in a collaborative task.

III. Methodology

The first phase of a three-phase study is currently underway to investigate the hesitation gesture as a means of handling uncertainties in a HRST. In the first phase hesitation gestures in a human-human shared-task (HHST) are identified and characterized quantitatively in terms of velocity, acceleration, and jerk. These characteristic motions are then implemented onto a robot arm such that the robot will exhibit hesitation gestures when encountering uncertainty or conflict in a HRST. In the second phase of this study, we will empirically determine whether the generated robot motions are also perceived by humans as representing hesitation. The third phase of this study will test the robot gestures' capacity to communicate its uncertainty to a user in a HRST.

A. Phase 1

Under the assumption that a human's hesitation gesture can be characterized in terms of the hand's linear velocity, acceleration and jerk, the first phase aims to quantitatively characterize hesitation gestures frequently observed in HHST environments. In this study, human subjects ($n_1 \cong 5$) are asked to engage in a shared-task with another person, with inertial sensors placed at various locations on one of the participants' dominant arm to collect linear and angular acceleration data. The task involves two people sorting a deck of cards together into appropriate bins according to various sorting rules. A pilot study showed this task to frequently cause hesitation gestures in human subjects. Video recordings obtained from the shared task will be broken down into discrete timeframe labels (A, B, C, etc.) and presented in an online survey in which another set of participants ($n_2 \cong 30$) will be asked to identify the instances where the sensor-equipped hands hesitated. Z-tests will be used to determine whether a given timeframe of a video contains a hesitation motion with statistical significance ($p<0.05$). The set of timeframes identified as containing a hesitation gesture will be the same timeframes of inertial sensor data used to characterize human hesitation gesture in terms of linear velocity, acceleration, and jerk. These characteristics will be used to generate robot hesitation gestures for a CRS robot arm.

B. Phase 2

We hypothesize that a robot motion having the same characteristics as that of human hesitation gestures will be perceived as hesitation. The second phase of this study will

statistically test this hypothesis. A video of the robot engaged in a shared task —analogous to the sorting task used in Phase 1—with a human will be broken down into discrete time-frame labels and subsequently shown in an online survey. The video will contain multiple instances of robot hesitation gestures. The participants of the survey ($n_3 \cong 30$) will be asked to identify frames in which the robot hesitated. The same statistical analysis carried out for Phase 1 will be used to determine which timeframes of the video were identified as containing a hesitation gesture. Results of this statistical analysis will be compared with the programmed occurrence of robot hesitation gestures. Any false positives and false negatives found will be tested for statistical significance.

C. Phase 3

Once a visually analogous hesitation motion of a robot is determined empirically in previous phases, we hypothesize that this motion can serve the same communicative function as that of a human's in conveying the robot's state of uncertainty. In Phase 3 of this study, participants ($n_4 \cong 20$) will be asked to engage in a HRST. Half of the participants will be engaged in a shared task with a robot that does not use hesitation gestures when encountering uncertainty or conflict, and the other half of the participants will be engaged in the same shared task with a robot that uses hesitation gestures. A post experiment survey will be conducted in order to study human perception of a robot in a HRST environment when the robot uses hesitation gestures. Likert-scale measurements will be collected to determine whether the HRST with robot hesitation gestures are perceived as friendlier than that of the HRST without hesitation motions.

In the future, we hope to implement a gesture recognition system such that the robot will not only be capable of exhibiting hesitation gestures, but also of recognizing a human's hesitation gesture. This bidirectional hesitation system can be used to mediate the decision of whether the robot or the human should yield the shared space or object. We hypothesize that this bidirectional communication via hesitation gestures will foster safer and friendlier interaction in HRST environments. The robot's actions will be seen as appropriate according to social norms and considerate of the user's internal state.

We believe that this work will aid in developing effective and appropriate methods of conveying a robot's state of uncertainty to a lay user during a collaborative human-robot task.

IV. ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the Institute for Computing, Information & Cognitive Systems.

REFERENCES

[1] W. Wallach and C. Allen, Moral Machines: Teaching Robots Right from Wrong. New York, NY: Oxford University Press, 2008, pp. 4.

[2] S. Schaal, "The New Robotics- towards human-centered machines," HFSP Journal Frontiers of Interdisciplinary Research in the Life Sciences, vol. 1, 2007, pp. 115-26.

[3] D. Kulic and E. A. Croft , "Real-time safety for human-robot interaction." Robotics and Autonomous Systems, vol. 54, no. 1, 2006, pp. 1-12.

[4] H.F.M. Van der Loos, "Ethics by Design: A Conceptual Approach to Personal and Service Robot Systems," ICRA Roboethics Workshop, Rome, Italy: IEEE, 2007.

[5] W. Ju and L. Takayama, "How People Interpret Automatic Door Movement as Gesture," International Journal of Design, vol. 3, no. 2, 2009, pp. 1-10.

[6] A. Green, K. Eklundh, B. Wrede, and S. Li, "Integrating Miscommunication Analysis in Natural Language Interface Design for a Service Robot," IEEE/RSJ International Conference on Intelligent Robots and Systems, 2006, pp. 4678-4683.

[7] C. Breazeal, "Social interactions in HRI: the robot view," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 34, no. 2, 2004, pp. 181-186.

[8] C. Breazeal, C. Kidd, A. Thomaz, G. Hoffman, and M. Berlin, "Effects of Nonverbal Communication on Efficiency and Robustness in Human-Robot Teamwork," 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005, pp. 383-388.

[9] H. F. M. Van der Loos and E. A. Croft, "Intrinsic, Multi-Modal Human Robot Communication," RO-MAN 2008 Workshop: Robots as Social Actors, Munich, Germany: IEEE 2008.

[10] D. B. Givens, The nonverbal dictionary of gestures, signs & body language cues, Spokane, Washington: Center for Nonverbal Studies Press, 2004.

[11] C. Suda and J. Call, "What Does an Intermediate Success Rate Mean? An Analysis of a Piagetian Liquid Conservation Task in the Great Apes," Cognition, vol. 99, no.1, 2006, pp. 53-71.

[12] J.F. Sousa-Poza and R. Rohrberg, "Body Movement in Relation to Type of Information (Person-and Nonperson- Oriented) and Cognitive Style (Field Dependence) 1," Human Communication Research, vol. 4, no. 1, 1977, pp. 19-29.

[13] P. Rober, "Some Hypotheses about Hesitations and their Nonverbal Expression in Family Therapy Practice," Journal of Family Therapy, vol. 24, 2002, pp. 187-204.

[14] G. Ball and J. Breese, "Relating Personality and Behavior: Posture and Gestures," Lecture Notes in Computer Science, Berlin / Heidelberg: Springer, 2000, pp. 196-203.

[15] H. Kim, S.S. Kwak, and M. Kim, "Personality Design of Sociable Robots by Control of Gesture Design Factors," Proc. 17th IEEE International Symposium on Robot and Human Interactive Communication, Munich: 2008, pp. 494-499.

[16] C.L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz and R. Alami, "A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction," Proc. IEEE International Workshop on Robots and Human Interactive Communication, 2005, pp. 371-377.

[17] P. Bremner, A.G. Pipe, M. Fraser, S. Subramanian and C. Melhuish, "Beat Gesture Generation Rules for Human-Robot Interaction,"18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan: 2009, pp. 1029-1034.

[18] Y. Muto, S. Takasugi, T. Yamamoto, and Y. Miyake, "Timing Control of Utterance and Gesture in Interaction between Human and Humanoid robot," 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan: 2009, pp. 1022-1028.

[19] C. Bethel, "Robots Without Faces: Non-Verbal Social Human-Robot Interaction," Ph.D. dissertation, University of South Florida, Tampa, Fl, USA, 2009.

[20] L. Takayama and C. Pantofaru, "Influences on proxemic behaviors in human-robot interaction," 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2009, pp. 5495-5502.

[21] T. Matsumaru, "Discrimination of Emotion from movement and Addition of Emotion in Movement to Improve Human-Coexistence Robot's Personal Affinity," 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan: 2009, pp. 387-394

[22] O. Mubin, A. Mahmud, and C. Bartneck, "TEMo-Chine : Tangible Emotion Machine,"Human-Computer Interaction - INTERACT 2007, 11th IFIP TC 13 International Conference, Rio de Janeiro, Brazil, 2007, Part 1, pp. 511-514.

Monitoring and Guiding User Attention and Intention in Human-Robot Interaction

Aaron St. Clair, Ross Mead, and Maja J. Matarić, *Senior Member, IEEE*

***Abstract*—A robot interacting with humans and attempting to generate effective social interaction and intervention behaviors benefits greatly from being able to understand and predict the underlying intentions of actions in context. Related work on collaborative discourse suggests that intention can be described in terms of either goal-directed task completion or communicative behavior directed to other collaboration partners. This paper describes early work on a generalizable framework for estimating the attentional space of a human interaction partner, providing context for grounding action in terms of intentions, and using this model to perform contextualized robotic intervention and ambiguity resolution. We describe an experiment aimed at applying and validating the framework in a simple collaborative human-robot interaction scenario involving deictic gestures.**

I. Introduction

One of the key challenges for the development of autonomous robots capable of effective interaction with humans is accurately detecting and reacting to human activity in a variety of interaction contexts. Our work is motivated by, but not limited to, socially assistive robotics (SAR) [1], an area of human-robot interaction focusing on helping people through social interaction. Preliminary work in USC's Interaction Lab has shown that SAR has the potential to benefit a diverse variety of target populations, including stroke patients, children with autism, and the elderly [2, 3].

Most existing systems for modeling human activity in interactive settings are closely tied to a specific control system, robotic platform, and task model. Adapting these systems to new task environments and robot embodiments is inherently difficult and often requires re-implementing the system from the bottom up. By imbuing robots with some form of social intelligence, we aim to unify common interaction mechanisms across a wide variety of populations and platforms. Toward these ends, we are investigating a general framework for monitoring the attentional space of a user, contextualizing specific user actions according to intent, and using this as a basis for formulating practical robot actions for intervening and communicating robot intention.

Manuscript received March 31, 2010. This work was supported in part by a National Science Foundation (NSF) Graduate Research Fellowship, as well as NSF grants CNS-0709296, IIS-0803565, and IIS-0713697.

A. St. Clair, R. Mead, and M. J. Matarić are with the University of Southern California, Los Angeles, CA 90089 USA. (phone: 213-740-6245; fax: 213-821-5696; email: {astclair, rossmead, mataric}@usc.edu).

II. User Monitoring

Human activity in social contexts is extremely complex. Determining the intended effect of human action as observed from sensor data is difficult; it requires filtering and segmenting the input serial data streams, assigning one of a number of possible explanations for any given action, and contextualizing the intended action in terms of the interaction.

Previous work on collaborative discourse theory and discourse coherence in a linguistic setting has emphasized constrained inference over salient information [4,5]. Drawing from this work, we model the interaction of the user and the robot as a simplified collaboration [4], in which the robot maintains a model of the user that is then used in both human action recognition and robot action planning. Breazeal *et al.* (2004) have used a related approach to structure collaborative learning, in which a robot learns a task model from guided human examples [6]. While our approach uses similar theoretical underpinnings, it is aimed at developing a general social communication framework that can be applied in many task environments, learned or otherwise.

A. Attention

User attention is modeled by constructing a probability distribution over salient world objects. Our current approach takes into account the user's position and head direction, extracted from camera and laser rangefinder sensors, as well as the relative saliency of world objects, to assign probability mass. For example, objects in the person's field of view are assigned relatively higher weights. The saliency map used can be specified *a priori* if the task domain is relatively static and well specified in advance, or it can be computed in real time using scene analysis [7].

B. Intention

Predicting human action and mapping it to underlying intention is a difficult problem since human activity is inherently complex. Neuroscience evidence suggests that humans accomplish this feat using mirror neurons to recognize actions while recruiting their own intentions for the recognized action and ascribing them to others [8]. Using a model of the current task and the estimated attentional space, we constrain the space of possible future actions and provide context to explain why a user might perform a particular action at a given time.

Our current focus is on deictic gestures—such as pointing, head orientation, and eye gaze—since they are well understood as a means of establishing joint attention, and are easily identified and physically grounded in terms of world objects. To compute possible target objects of a pointing gesture with respect to the user, we can utilize a Bayesian approach to combine an error model of human pointing and the attentional distribution as a prior. We are investigating methods for recognizing attention and action stemming from more complex intentions and distinguishing that are task-oriented actions, such as reaching, from communication-oriented actions, such as pointing and other social gestures.

III. User Intervention

By monitoring user intentions and anticipating their effect on the success of an individual or collaborative task, a robot may determine that it is appropriate for it to intervene. Such an intervention may be deemed necessary to improve task performance or to prevent undesirable actions from being taken by the user. Directing user attention and intention must be done as clearly as possible to maintain a successful interaction between the robot and user. It is therefore crucial that potential ambiguity be minimized or resolved.

A. Intervention

In this preliminary work, the robot plans and executes an intervention strategy over possible proxemic and deictic actions. *Proxemics* here refers to the manipulation of robot position and orientation with respect to the human [9]. The robot must situate itself in the appropriate "social space" to maximize the effectiveness of subsequent communicative actions. Once the robot has positioned itself, it utilizes *deictic gestures*—such as pointing, head orientation, and eye gaze—to focus the attention of the user to a particular object or region, thus attempting to establish joint attention [10]. Intent is then communicated by exploiting the theory of perceived *affordances*, which suggests how an object may be interacted with [11]. This reliance on affordances constrains the interaction to simple tasks; however, in future work, we will investigate more complex forms of representation and communication of intent [12], and extend our probabilistic framework to consider McNeill's four categories of discourse gestures (iconic, metaphoric, deictic, and beat) [13].

B. Ambiguity Resolution

In the ideal case, the appropriate application of social distance and deictic gestures would result in a clear user interpretation of the task objective and, thus, a successful intervention; however, in the real world, such communication is often noisy and potentially ambiguous. To resolve such ambiguity, the robot engages in perspective-taking, considering the viewpoint of the human observer, as well as previous user activity. We utilize a naïve Bayes approach to estimate the clarity of a human's interpretation of potential robot actions over the attentional space. We then select a robot intervention strategy by applying gradient decent to find a global minimum with regard to ambiguity.

IV. Implementation

We are in the process of collecting human interpretation data based on interactions with a physical robot. From this, we can produce a probabilistic model of error in human perception of robot deictic gestures. This model will then be validated in a collaborative task to demonstrate the efficacy of robot intervention and ambiguity resolution strategies and attention and intention monitoring with a human user.

A. Robot Platform

The system is being implemented on the Bandit III robot platform available in the Interaction Lab, shown in Fig. 1. Bandit is an upper-torso humanoid robot with 17 degrees of freedom: 7 in each arm (shoulder forward and backward, shoulder in and out, elbow tilt, elbow twist, wrist twist, grabber open and close; left and right arms), 2 in the head (pan and tilt), 2 in the lips (upper and lower), and 1 in the eyebrows. These degrees of freedom allow the robot to be highly expressive through individual and combined motions of the head, face, and arms. An extensive gesture and facial expression library has been developed to enhance the interactive experience. The robot is closer to human-scale than many other humanoid platforms; mounted atop a Pioneer P2 base, the entire robot stands one meter tall, making it an adequate choice for robot interaction. An overhead camera and on-board laser rangefinder facilitate human and robot pose tracking.

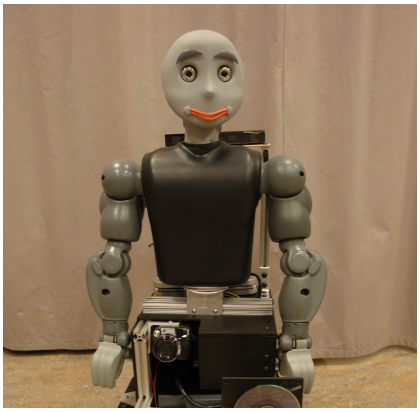


Fig. 1. The Bandit upper-torso humanoid robot platform

B. Experiment Design

We are investigating a concrete application of this framework within the realm of deictic gesture. The experimental design is a two-phased approach aimed at producing an empirical error model of both human gesture perception accuracy and robot gestural accuracy, and then applying these models using our attention, intention, and ambiguity resolution framework to allow a robot to engage in a simple collaborative task with a human partner.

1) Building perceptual models: We have begun preliminary experiments aimed at building an error model for human perception based on different robot pointing modalities, including head, arm, and combined head and arm gestures. Each gesture's accuracy is evaluated in

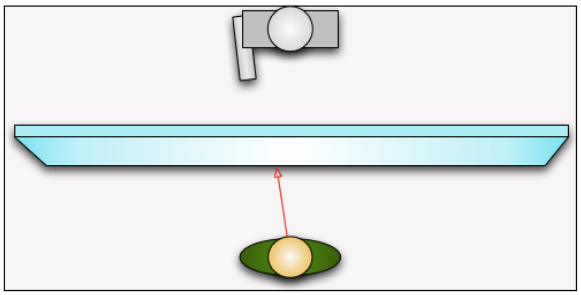


Fig. 2. The robot makes a deictic gesture on one side of the screen, and the user indicates a perceived point.

experiments with human participants, in which the angle, distance and point location are systematically varied. Target points are assumed to be on an approximately 2.5 x 3.5 m (8 x 12 ft) transparent acrylic screen. The participant is seated on one side of this screen, opposite the robot, as illustrated in Fig. 2. As the robot gestures to each point, the participant uses a laser pointer to indicate his perception of the target location on the screen. An experimenter then marks these points with a fiducial marker, and they are recorded using an upward-pointing laser rangefinder, yielding measurements accurate to within centimeters.

In the first iteration of this experiment all users were given a random set of points on a regular grid and we aimed to use within subject comparison to determine how error responds to changes in the state variables. After conducting an analysis of variance from data collected from 11 participants we determined that errors were generally on the order of 30-60cm (1-2 feet), but were unable to draw deeper conclusions due to high variance in the sampled distributions. To address these issues we have redesigned the experiment to perform a between subjects comparison, with more participants. This will make use of an interspersed "calibration point," within the other randomly distributed points presented to each user, to ensure that within-user accuracy is consistent. From this output, we construct an error model parameterized by human-robot-point locations and angles. During this experiment, we are also monitoring the head orientation of the participant in order to empirically determine if head direction can be used to model attention.

2) Validating attention, intention and intervention: Using this perceptual error model, we will conduct a further experiment to validate the estimated attentions and intentions within the context of a collaborative game-playing scenario. The game involves a robot indicating to the user a series of targets within a cluttered office environment; the user must then visit these targets in a specified order; this task is similar to, but less constrained than, that of our previous work [14], and was chosen specifically for comparison and analysis. The error model of human perception will be used to determine the position and orientation from which the robot should point to a target to ensure that the gesture is specified in a minimally (or maximally, for testing purposes) ambiguous manner. We are also investigating the use of a similar intention model to determine, at any point in time, the most likely intended target of a user, allowing the robot to intervene by redirecting the user, if necessary, to correct potential errors.

REFERENCES

[1] D. Feil-Seifer and M. J. Matarić, "Defining socially assistive robotics," *Proceedings of the International Conference on Rehabilitation Robotics (ICORR-05)*, Chicago, Illinois, 2005.

[2] A. Tapus, M. J. Matarić, and B. Scassellati, "The grand challenges in socially assistive robotics," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, pp. 35-42, 2007.

[3] D. J. Feil-Seifer and M. J. Matarić, "Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders," *11th International Symposium on Experimental Robotics 2008*, vol. 54, pp. 201-210, Athens, Greece, July 2008.

[4] B. Grosz and S. Kraus, "The evolution of SharedPlans," *Foundations of Rational Agency*, vol. 14, pp. 227–262, 1999.

[5] A. Lascarides and M. Stone, "Discourse coherence and gesture interpretation," *Gesture*, vol. 9, no. 2, pp. 147–180, 2009.

[6] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Chilongo, "Tutelage and collaboration for humanoid robots," *International Journal of Humanoid Robotics*, vol. 1, no. 2, pp. 315–348, 2004.

[7] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[8] S. Blakemore and J. Decety, "From the perception of action to the understanding of intention," Nature Reviews Neuroscience, vol. 2, no. 8, pp. 561–567, 2001.

[9] M. L. Knapp and J. A. Hall, *Nonverbal communication in human interaction*, 7th ed., Boston, MA: Wadsworth Publishing, 2009.

[10] M. M. Louwerse and A. Bangerter "Focusing attention with deictic gestures and linguistic expressions", *Proceedings of the XXVII Annual Conference of the Cognitive Science Society (CogSci 2005)*, Stresa, Italy, July 21–23 2005.

[11] D. A. Norman, *The design of everyday things*, New York: Basic Books, 2002.

[12] R. Mead, J. B. Weinberg, and M. J. Matarić, "An ontology-based multimodal communication system for human-robot interaction in socially assistive domains", *IEEE ICRA2010 Workshop on Multimodal Human-Robot Interfaces*, Anchorage, AK, May 2010.

[13] D. McNeill, *Gesture and thought*, Chicago: University of Chicago Press, 2005

[14] R. Mead and M. J. Matarić, "The power of suggestion: teaching sequences through assistive robot motions," *Proceedings of The 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI-09)*, San Diego, CA, pp. 317-318, March 2009.

Generating Multi-Modal Robot Behavior Based on a Virtual Agent Framework

Maha Salem, Stefan Kopp, Ipke Wachsmuth, Frank Joublin

***Abstract*— One of the crucial steps in the attempt to build sociable, communicative humanoid robots is to endow them with expressive non-verbal behaviors along with speech. One such behavior is gesture, frequently used by human speakers to emphasize, supplement, or even complement what they express in speech. The generation of speech-accompanying robot gesture together with an evaluation of the effects of this multi-modal behavior is still largely unexplored. We present an approach to systematically address this issue by enabling the humanoid Honda robot to flexibly produce synthetic speech and expressive gesture from conceptual representations at run-time, while not being limited to a predefined repertoire of motor actions in this. Since this research challenge has already been tackled in various ways within the domain of virtual conversational agents, we build upon experiences gained with speech-gesture production models for virtual humans.**

I. INTRODUCTION

Humanoid robot companions that are intended to engage in natural and fluent human-robot interaction in rich environmental settings must be able to produce speech-accompanying, non-verbal behaviors. Forming an integral part of human communication, hand and arm gestures are primary candidates for extending the communicative capabilities of social robots. This, however, poses a number of research challenges, especially with regard to a motor control for arbitrary, expressive hand-arm movement and its coordination with other interaction modalities such as speech. The generation of co-verbal gestures for artificial humanoid bodies demands a high degree of control and flexibility concerning shape and time properties of the gesture, while ensuring a natural appearance of the movement. Ideally, if such non-verbal behaviors are to be realized, they have to be derived from conceptual, to-be-communicated information.

Since the challenge of multi-modal behavior realization has already been explored in various ways within the domain of virtual conversational agents, our approach builds upon the experiences gained from the development of a speech and gesture production model used for the virtual human *Max* [2]. Being one of the most sophisticated multi-modal schedulers, the Articulated Communicator Engine (ACE) has replaced the use of lexicons of canned behaviors with an on-the-spot production of flexibly planned behavior representations. Employing it as an underlying action generation architecture for the Honda humanoid robot, ACE draws upon a tight, bi-directional coupling of the robots perceptuo-motor system with multi-modal scheduling via both efferent control signals and afferent feedback.

The work described is supported by the Honda Research Institute Europe.

M. Salem is at the Research Institute for Cognition and Robotics, Bielefeld, Germany msalem@cor-lab.uni-bielefeld.de

S. Kopp is at the Sociable Agents Group, Bielefeld University, Germany skopp@techfak.uni-bielefeld.de

I. Wachsmuth is at the Artificial Intelligence Group, Bielefeld University, Germany ipke@techfak.uni-bielefeld.de

F. Joublin is at the Honda Research Institute Europe, Offenbach, Germany frank.joublin@honda-ri.de

II. SPEECH-GESTURE PRODUCTION MODEL FOR A HUMANOID ROBOT

Within the ACE framework, there are two different ways to describe gesture representations using the XML-based Multi-modal Utterance Representation Markup Language (MURML [3]). Firstly, verbal utterances in combination with co-verbal gestures can be specified with feature-based descriptions. In such MURML utterances, the outer form features of a gesture (i.e., the posture designated for the gesture stroke) are explicitly described. Their affiliation to dedicated linguistic elements is determined by matching time identifiers. Fig. 1 illustrates an example of a feature-based MURML specification that can be used as input for speech-gesture production. Secondly, gestures can be specified as keyframe animations whereby each keyframe specifies a 'key posture', a part of the overall gesture movement pattern describing the current state of each joint. Speed information for the interpolation between every two key postures and the corresponding affiliation to parts of speech is obtained from assigned time identifiers. Keyframe animations in ACE can be either defined manually or, alternatively, derived from motion capturing data from a human demonstrator, allowing the animation of virtual agents in real-time.

A. On-line Scheduling of Multi-Modal Utterances

In a given multi-modal utterance, each intonation phrase together with a co-expressive gesture phrase represents a

```
<definition><utterance>
 <specification>
  The bathroom is <time id="t1"/> over there. <time id="t2">
 </specification>
 <behaviorspec>
  <gesture id="gesture_1" scope="hand">
   <affiliate onset="t1" end="t2" focus="there"/>
    <constraints>
     <parallel>
      <static slot="HandShape" value=" BSflat (FBround all o)"/>
      <static slot="ExtFingerOrientation" value="DirA"/>
      <static slot="PalmOrientation" value="DirR"/>
      <static slot="HandLocation" value="LocShoulder LocCenterLeft LocStretched"/>
     </parallel>
    </constraints>
  </gesture>
 </behaviorspec>
</utterance></definition>
```

Fig. 1. Example of a feature-based MURML specification for multi-modal utterances.

single idea unit which is referred to as a *chunk* of speech-gesture production [2]. Incremental production of successive coherent chunks is realized by processing each chunk on a separate 'blackboard' running through a sequence of states (Fig. 2). Speech-gesture synchronization within a chunk is

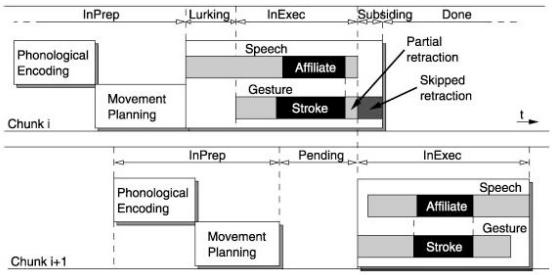


Fig. 2. Blackboards running through a sequence of processing states for incremental production of multi-modal chunks [2].

achieved on-line by the ACE engine by adapting the gesture to structure and timing of speech. To do this, the ACE scheduler retrieves timing information about the synthetic speech at the millisecond level and defines the start and the end of the gesture stroke accordingly. These temporal constraints are automatically propagated down to each single gesture component. A more detailed overview of the internal planning process within ACE can be found in [2].

B. Speech Synthesis

Spoken utterances are generated using the open source text-to-speech synthesis system MARY (Modular Architecture for Research on speech sYnthesis) [4]. Its main features are a modular design and an XML-based internal data representation. Several languages are supported including English and German. For further details on MARY see [4].

C. Robot Control Architecture

In order to enable the humanoid robot to flexibly produce speech and co-verbal gesture at run-time, a robot control architecture is required which combines conceptual representation and planning provided by ACE with motor control primitives for speech and arm movements for the robot. This endeavor poses a number of interesting challenges including a failure to adequately account for certain physical properties – motor states, maximum velocity, strict self collision avoidance, variation in DOFs, etc. This is in light of ACE being originally designed for a virtual rather than physical platform. Hence, when transferring the ACE framework to the physical robot these challenged must be met.

Since gesture generation with ACE is based on external form features as annotated in the MURML specification, our robot control architecture suggests that arm movement trajectories are described directly in task space. The information obtained at the task space level including wrist orientation and designated hand shape is forwarded to the robot motion control module which instantiates the actual robot movement. Inverse kinematics (IK) of the arm is then solved on the velocity level using the whole body motion (WBM) controller framework [1]. The WBM framework allows to control all DOFs of the Honda humanoid robot based on given end-effector targets, providing a flexible method to control upper body movement by only specifying relevant task dimensions selectively in real-time, yet, while generating smooth and natural movement. Redundancies are optimized with regard to joint limit avoidance and self-collision avoidance. For more details on WBM control for the Honda humanoid research robot see [1].

After IK has been solved for the internal body model provided for WBM control, the joint space description of the designated trajectory is applied to the physical robot. A bi-directional interface using both efferent actuator control signals and afferent sensory feedback monitors possible deviations of actual robot motor states from the kinematic body model provided by ACE. It is realized by a feedback loop that updates the internal model of the robot in the WBM controller as well as the kinematic body model coupled to ACE at a sample rate *r*. Fig. 3 illustrates our robot control architecture embedding the ACE framework.

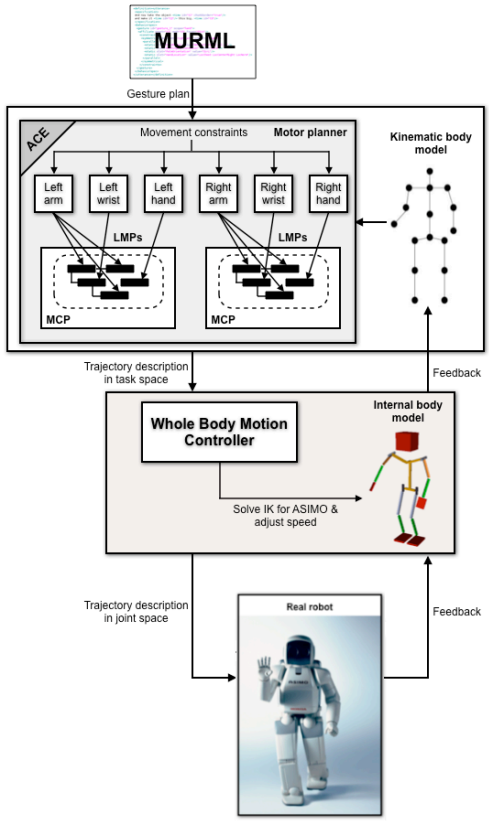


Fig. 3. Robot control architecture for the generation of gesture behavior.

III. Results and Discussion

Results were produced in a feed-forward manner whereby commands indicating the wrist position and hand orientation of the ACE body model were constantly transmitted to the robot at a sample rate of 20 frames per second. IK was solved using the provided whole body motion controller.

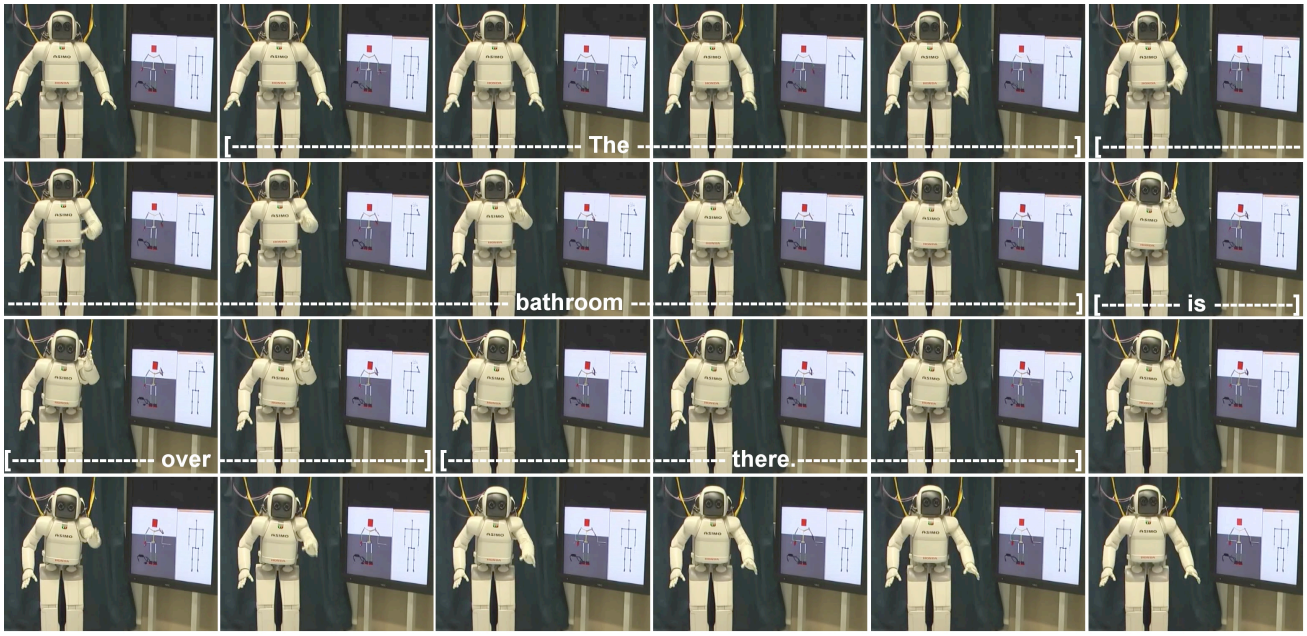


Fig. 4. Example of a multi-modal utterance realized in the current framework, allowing comparison of the physical robot, the internal robot body model and the ACE kinematic body model (left to right, top-down, sampled every four frames (0.16sec)).

Speech output was synthesized using the MARY text-to-speech system based on the multi-modal utterance scheduler in ACE. Fig. 4 illustrates the multi-modal output generated in our current framework using the MURML utterance presented in Fig. 1. The robot is shown next to a panel which displays the current state of the internal robot body model and ACE kinematic body model, respectively, at each time step. In addition, speech output is transcribed to illustrate the words spanning different segments of the gesture movement sequence, indicating temporal synchrony achieved between the two modalities. It is revealed that the physical robot is able to perform a generated gesture fairly accurately but with some inertial delay compared to the internal ACE model. Despite the general limitation in motion speed, these findings substantiate the feasibility of the proposed approach. Arbitrary MURML-based speech-gesture representations can be realized using the current framework. Synchronization of speech and gesture, however, does not appear to be optimal yet. Although Fig. 4 suggests acceptable temporal synchrony between both output modalities, tests using long sentences in speech revealed that movement generation tends to lag behind spoken language output. Consequently, we need to explore ways to handle the difference in time required by the robot's physically constrained body in comparison to the kinematic body model in ACE. Our idea is to tackle this challenge by extending the cross-modal adaptation mechanisms provided by ACE with a more flexible multi-modal utterance scheduler which will allow for a finer mutual adaptation between robot gesture and speech.

IV. Conclusion and Future Work

We presented a robot control architecture which enables the Honda humanoid research robot to generate gestures and synchronized speech at run-time. Meeting strict temporal synchrony constraints will present a main challenge to our framework in the future. Evidently, the generation of finely synchronized multi-modal utterances proves to be more demanding when realized on a robot with a physically constrained body than for an animated virtual agent. To tackle this new dimension of requirements, however, the cross-modal adaptation mechanisms applied in ACE have to be extended to allow for a finer mutual adaptation between robot gesture and speech.

Our results help to shed light on conceptual motorics in robotic agents. Essentially, they substantiate the feasibility of our approach while pointing out the direction for our future research. Once our robot control architecture has been extended to account for a finer synchronization of gesture and speech, it will be assessed in human-robot interaction studies, providing new insights into human perception and understanding of gestural machine behaviors and how these can be used to design more natural communication in robots.

References

[1] M. Gienger, H. Janßen, and S. Goerick. Task-oriented whole body motion for humanoid robots. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, Tsukuba, Japan, 2005.

[2] S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.

[3] A. Kranstedt, S. Kopp, and I. Wachsmuth. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents - let's specify and evaluate them*, Bologna, Italy, July 2002.

[4] M. Schröder and Jürgen Trouvain. The german text-to-speech synthesis system mary: A tool for research, development and teaching. In *International Journal of Speech Technology*, pages 365–377, 2003.

Robot, tell me what you know about...?: Expressing robot's knowledge through interaction

Raquel Ros, E. Akin Sisbot, Séverin Lemaignan, Amit Pandey and Rachid Alami
CNRS - LAAS, 7 avenue du Colonel Roche, F-31077 Toulouse, France
Université de Toulouse, UPS, INSA, INP, ISAE, LAAS, F-31077 Toulouse, France
Email: {rrosespi, slemaign, easisbot, akpandey, rachid}@laas.fr

***Abstract*— Explicitly showing the robot's knowledge about the states of the world and the agents' capabilities in such states is essential in human robot interaction. This way, the human partner can better understand the robot's intentions and beliefs in order to provide missing information that may eventually improve the interaction. We present our current approach for modeling the robot's knowledge from a symbolic point of view based on an ontology. This knowledge is fed by two sources: direct interaction with the human, and geometric reasoning. We present an interactive task scenario where we exploit the robot's knowledge to interact with the human while showing its internal geometric reasoning when possible.**

I. Introduction

"Why is this robot doing this now?!" This is the typical question that at some point a user asks herself when interacting with a robot. And most probably, not only naive users, but also robot designers when working with their robots. Understanding and knowing the robot's internal knowledge and reasoning states is fundamental to improve any type of interaction. Feedback is specially crucial when a problem occurs, or when the robot makes an unexpected decision. Ideally, this feedback should be given through a natural dialogue where the robot explains its decisions and actions. In order to have robots capable of reasoning on their own internal states to naturally communicate with their human partners, several supportive mechanisms should be considered.

In this work we introduce several mechanisms and their connection applied to a face-to-face interactive task. In this task the human asks the robot about its knowledge on objects in the environment and about its reasoning on the agents' abilities in the world. Thus, we present an approach for modeling the robot's knowledge based on an ontology (Sec. II) and a geometric reasoner that transforms geometric world information into symbolic descriptions (Sec. III). A decisional reasoner interprets the human query (entered through keyboard) in order to identify the referred object (Sec. IV) and then queries the robot's knowledge about it to answer the human. Human queries are limited to a fix vocabulary and a specific format (interpretation of natural language is out of the scope of this work). The reply is at least given by spoken language (and written on the screen), and if possible, by visual feedback from the robot's internal 3D environment model.

This research was supported by a Marie Curie Intra European Fellowship and the European Community's Information and Communication Technologies within the 7th European Community Framework Programme under grant agreements no. [220368], ARBI and [215805] CHRIS projects.

II. Knowledge Representation

We believe that the knowledge model of a robot should include a comprehensive model of the roles, relationships and context of objects in the environment, as well as beliefs and intentions of other agents. Moreover, this understanding must rely on a formal encoding that requires high expressivity while remaining well suited for machine processing in order to be used by the robot.

We thus propose the use of ORO (the "OpenRobot Ontology" server), a central knowledge repository that stores, manages, processes and exposes knowledge for the robot from a symbolic point of view. It internally relies on RDF-derivate OWL Description Logics to formally represent statements on the world as triples `<subject> <predicate> <object>`. It uses two open-source libraries: *Jena* for storage and manipulation of statements and *Pellet* first-order logic reasoner to classify, apply rules and compute inferences on the knowledge base [1].

ORO defines an initial *upper* ontology for human-aware robotics called *OpenRobots Commonsense Ontology*. This initial ontology contains a set of concepts, relationships between concepts and rules and defines the "cultural background" of the robot, i.e. the a priori known concepts. Currently, this commonsense knowledge is focused on the requirement of human-robot interactions in everyday environments, but contains as well generic concepts like `thing`, `object`, `location` and relationships between those. The common-sense ontology design relies heavily on the standard OpenCyc upper ontology for the concepts naming, thus ensuring a good compatibility with other knowledge bases. Figure 1 illustrates a simple example with some concepts.

Besides simply storing and reasoning about knowledge, ORO offers several useful features for human-robot interaction. One advantage offered by the ORO architecture is that independent cognitive models for each agent can be maintained. When the robot interacts with a new agent, a separate RDF triple storage is created to store the robot's knowledge about the agent's perception. For instance, in the case of perspective taking, we compute the visibility and spatial information about the world from each agent point of view, and store it in their own cognitive models. Having

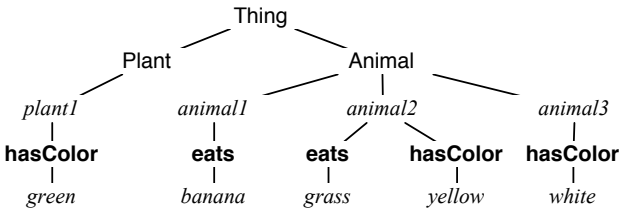


Fig. 1. Ontology example. Names with first capital letter correspond to classes; bold names, to properties; and italic names, to instances.

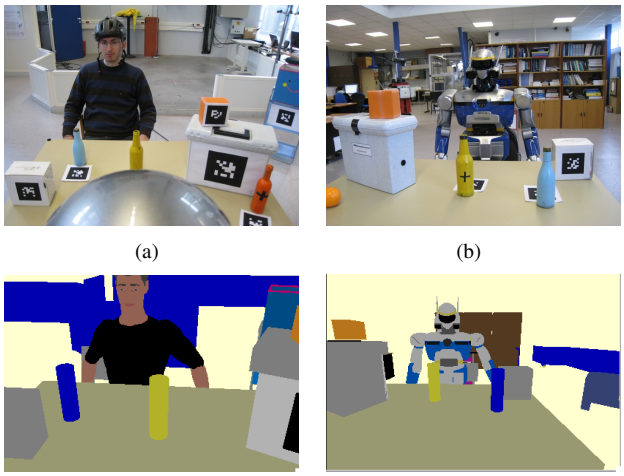


Fig. 2. Visual perspective taking for each agent in the scenario.

separate cognitive models allows us to store and reason on different models of the world.

III. Geometrical Reasoning

This section describes different reasoning mechanisms to provide an abstraction layer to the decisional layer on top of the geometrical description of the environment.

To model the environment we use the software platform Move3D [2]. The kinematic structures of the human and the robot, as well as their positions and objects' positions are integrated into this platform to maintain a coherent model of the real environment. It also allows us to view the visual perspective of the agents in the world by modeling their visual sensors (eyes for humans, cameras for robots).

We divide the geometrical reasoning mechanisms in two groups: perspective taking descriptors and symbolic location descriptors. The first set corresponds to information obtained when reasoning from an agent point of view, while the second one corresponds to global descriptors independent of the agents in the environment. All this information is stored in the ontology, which in turn may infer additional information as we explain next. Moreover, the information concerning specific agents, i.e. perspective taking descriptors, is stored in each agent's cognitive kernel in ORO allowing the decisional level to reason about each agents' beliefs about the world.

A. Perspective Taking Descriptors

1) Visibility: Visual perspective taking refers to the ability for visually perceiving the environment from other's point of view. This ability allows us to identify objects or regions that are visible/hidden for/from others, which is essential for referring to things when interacting with others. For example, if the robot refers to an object, it should try to ensure that the human is able to see it in order to facilitate interaction. On the contrary, if the human refers to an object, based on the context, she could refer to a visible one (e.g. "take this ball") or to an invisible one (e.g. "find the ball").

We are currently able to compute "visibility" from an agent point of view for objects in the environment [3] and zones or regions around the agent [4]. An object or a region is visible for an agent if, while performing a minimum effort (i.e. only turning the head or standing, if possible), the object or region are within the agent's field of view and there are no occlusions in between.

2) Spatial: Spatial perspective taking refers to the qualitative spatial location of objects (or agents) with respect to a frame of reference (eg. the keys on my left). Based on this frame, the spatial description of an object varies. Humans mix perspectives frequently during interaction [5], i.e. they do not maintain a consistent perspective through a conversation. Therefore, the robot has to be able to understand and compute descriptions of objects based on different frames of reference to follow the interaction with its human partner.

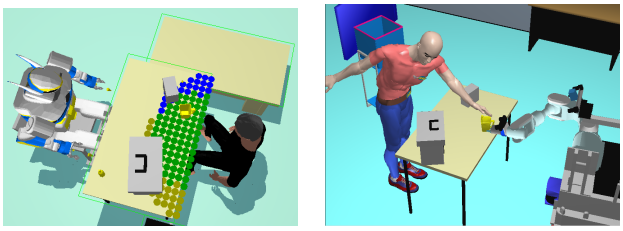
In this work, we use two types of the frames of reference: egocentric (from the robot perspective) and addressee-centered (from the human perspective). Thus, given an object and the referent, we can compute the spatial locations by dividing the space around the referent into n regions based on arbitrary angle values relative to the referent orientation. For example, for $n = 4$ we would have the space divided into front, left, right and back. Further subdivisions can be computed if we would like to represent distinctions among distances, e.g. near and far.

3) Reachability: An object or a region is reachable if there is a collision free posture for the agent where the end-effector is at the center of the object or region with a given tolerance. A valid posture includes moving the upper-body or standing, if possible.

This ability allows the robot to estimate the agent's capacity to reach an object, which is fundamental for task planning. For example, if the human asks the robot to give her an object, the robot must compute a transfer point where the human will be able to get the object. Figure 3 illustrates the reasoning results for reaching regions and an object.

B. Symbolic Location Descriptors

Symbolic location descriptors allow the robot to compute spatial relations between objects in the environment. The system infers symbolic relations between objects from its 3D geometric world representation. In this work we propose the use of three basic symbolic relations between each pair of objects. However, their inverse relations can be automatically computed at the symbolic level, i.e. through inference



(a) (b)

Fig. 3. (a) Reachable points from the human perspective when bending: yellow, blue and green points correspond to left hand, right hand and both hands respectively. (b) Human and robot posture for reaching the cup.

based on OpenRobots Commonsense Ontology, enlarging the symbolic descriptions knowledge easily.

- *IsIn*: indicates if an object (or an agent) is inside of another object. Its inverse relation corresponds to *Contains*.
- *IsOn*: indicates if an object (or an agent) is placed on top of another object. Its inverse relation is *IsUnder*[1].
- *IsNextTo*: tests if an object (or an agent) is next to another object. It has no inverse relation, but symmetric.

IV. Finding the Referent

Given partial (or complete) description of an object (list of attribute-value pairs), the robot is able to identify the referred object the following way. First it obtains all objects that fulfill the initial description. Based on the result it either succeeds (obtains one single object), fails (no object with that description could be found) or obtains several objects. In this latter case, a new descriptor is added to the initial description and the process starts over again. Failure occurs when the description does not match any object from the robot's knowledge. This could be because the robot's knowledge is incomplete (the human refers to an unknown descriptor or descriptor value) or due to inconsistent information (human's and robot's beliefs are different).

In order to automatically add a new descriptor (attribute-value pair), the reasoning engine must find the best discriminant for the current list of objects being evaluated. If found, the robot asks the human for its value. Discriminants are descriptors that allow a maximum discrimination among a set of individuals (eg. color, type, location, etc.). We distinguish two types of discriminants. *Complete* discriminants are those attributes (or properties) whose values can uniquely identify those individuals. However, they are not always available. First, because two or more individuals may share the same value, and second, because not all individuals may share the same properties. Thus, *partial* discriminants are those properties that "better" split the set of individuals in different subsets based on some criteria. In the task we propose in this work we only make use of complete discriminants, although partial discriminants may be useful for other tasks, as the Spy Game introduced in [6].

[1]We consider that there is a physical contact between both objects, although the English definition of under does not necessarily imply it.

The algorithm to find discriminants has the following steps (to better follow it we show an example corresponding to the ontology shown in Fig. 1). We search a discriminant for the following individuals: $plant_1, animal_2$ and $animal_3$. First we obtain the direct properties for all the individuals, i.e. we do not consider all the hierarchy of properties. In the example, we only take the most direct class for $plant_1$, i.e. the class `plant` (and not the class `thing`). Next, we compute the number of individuals per property and the number of different values for that property. If there is more than one different value for the property (in other words, if not all individuals have the same value), then we consider that property as a potential discriminant. Finally, we sort the list of potential properties following two criteria: number of individual occurrences (i.e. the most individuals are covered by that property, the better) and values occurrences (i.e. the more distinct values, the better). The best discriminant corresponds to the first element of the sorted list. If several properties are equal, return all of them. In our example, the algorithm would return the property `hasColor`.

V. Application

We have designed an interactive task that exploits the robot's knowledge while using the different mechanisms presented in this work. The scenario for the task consists in a face-to-face interaction around a table with objects. The human may ask the robot the following questions:

1) Where is the object_description?: The robot indicates the location of the object based on spatial perspective taking and symbolic location descriptors.

2) Is the object_description visible?: The robot computes the visibility of the object from both agents' perspectives (robot and human) and indicates whether the object is visible or not. If it is, it also indicates if it is directly visible (within the agent's current FOV) or if the object is visible by turning the head (out of FOV). The view of the agents is displayed in the screen at the same time (Figures 2c and 2d).

3) Is the object_description reachable?: The robot computes the object's reachability from both agents' perspectives (robot and human) and indicates whether the object is reachable or not. The screen displays the movement of the agent for reaching the computed posture (Figure 3b).

From the human query, the decisional reasoner extracts the description of the referred object. The description can be either the id (eg. YELLOW_BOTTLE) or a set of attributes about the object (eg. yellow bottle). In the first case, if the id corresponds to a known object, the robot can directly answer the question. On the contrary, if the object is unknown, then a learning phase takes place, where the human describes recursively the type of the object until a known type is reached. Figure 4 shows an example of the learning process where the robot asks for the type of object until recognizing a known type. In the second case, where attributes describing the object are given, the reasoning engine has to figure out the id of the object. However, the information provided by the human may be incomplete or ambiguous. Therefore, the robot must ask the human additional information until

human	(Shows an object to the robot).
robot	(Identifies an unknown object.) What's the name of the object?
human	mist
robot	What kind of object is mist?
human	laptop
robot	What is a laptop?
human	computer
robot	What is a computer?
human	object
robot	Now I know what mist is.

Fig. 4. Learning process example where the robot learns that mist is a computer.

human	Where is the bottle?
robot	*(Available object ids: BLUE_BOTTLE, YELLOW_BOTTLE and ORANGE_BOTTLE.)* *(Complete discriminants:* hasColor) Which color is the object?
human	Yellow.
robot	The YELLOW_BOTTLE is located: - from my perspective: FRONT - from your perspective: LEFT - on HRP2TABLE.
human	Is the orange bottle visible?
robot	*(Available object ids: ORANGE_BOTTLE)* The ORANGE_BOTTLE is visible for me (out of FOV) and not visible for you.
human	Is the white box on my left reachable?
robot	*(Available object ids: SPACENAVBOX)* The SPACENAVBOX is reachable for me and reachable for you.

Fig. 5. Interactive task example.

identifying the object [7], [6]. Figure 5 illustrates different interaction scripts where the robot finds out the referred object and answers the human query.

VI. Related Work

Few work has been developed within the robotics field regarding perspective taking reasoning, which is fundamental for enhancing social interaction ([8],[9],[10]). Trafton et al. use both visual and spatial perspective taking for finding out the referent indicated by a human partner [10]. In [11], they also design a robot that is able to play hide and seek a child does. The strategy they model in the robot is to find those places that are not visible for the human partner. Berlin et al. [12] present their work focused on using visual perspective taking skills for learning from a human teacher. A teacher classifies objects in a given way. The robot then learns the classification function based on the teacher's visual perception of the world. Johnson and Demiris [13] apply visual perspective taking for action recognition. In their work, a robot who has complete visual access of the environment observes another robot with partial access performing a task. The first robot can recognize the task performed by the second robot because it is able to reason about its partial perception. The most significant work for computing reachability has been introduced by Zacharias et al. [14], but only from the robot point of view and not the human, as we do in our work.

The novelty of our work is that we combine (1) different geometric reasoning mechanisms from both, human and robot perspective, which allows us to reason about the agent's capacities with (2) a symbolic knowledge representation, which allows us to reason about the agents' knowledge about the state of the world.

VII. Conclusions

We have presented a set of mechanisms to ease interaction between humans and robots while communicating the robot's internal knowledge about the world. More precisely, we have introduced a model for knowledge representation along with a geometric reasoning engine that provides symbolic descriptions of geometric relations, as well as agent's abilities. The overall system is completely platform independent and has been integrated in two different platforms.

References

[1] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz, "Oro, a knowledge management module for cognitive architectures in robotics," 2010, submitted to IROS.

[2] T. Siméon, J.-P. Laumond, and F. Lamiraux, "Move3D: A generic platform for path planning," in *IEEE International Symposium on Assembly and Task Planning, ISATP*, 2001, pp. 25–30.

[3] L. F. Marin-Urias, E. A. Sisbot, and R. Alami, "Geometric tools for perspective taking for human-robot interaction," in *7th International Conference on Artificial Intelligence*, 2008.

[4] A. K. Pandey and R. Alami, "Mightability maps: A perceptual level decisional framework for co-operative and competitive human-robot interaction," 2010, submitted to IROS.

[5] B. Tversky, P. Lee, and S. Mainwaring, "Why do speakers mix perspectives?" *Spatial Cognition and Computation*, vol. 1, 1999.

[6] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Which One? Grounding the Referent Based on Efficient Human-Robot Interaction," 2010, submitted to RO-MAN.

[7] R. Ros, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Solving ambiguities with perspective taking," in *In Proceedings of International Conference on Human-Robot Interaction*, 2010, pp. 181–182.

[8] B. Tversky and B. M. Hard, "Embodied and disembodied cognition: Spatial perspective-taking," *Cognition*, vol. 110, pp. 124–129, 2009.

[9] S. Wu and B. Keysar, "The effect of culture on perspective taking," *Psychological Science*, vol. 18, no. 7, pp. 600–606, 2007.

[10] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Transactions on systems, man and cybernetics - Part A: Systems and Humans*, vol. 35, no. 4, pp. 460–470, 2005.

[11] J. G. Trafton, A. C. Schultz, , D. Perznowski, M. D. Bugajska, W. Adams, N. L. Cassimatis, and D. P. Brock, "Children and robots learning to play hide and seek," in *Proceedings of the 1st Conference on Human-Robot Interaction*. ACM, 2006, pp. 242–249.

[12] M. Berlin, J. Gray, A. L. Thomaz, and C. Breazeal, "Perspective taking: An organizing principle for learning in human-robot interaction," in *In Proceedings of AAAI*, 2006.

[13] M. Johnson and Y. Demiris, "Perceptual perspective taking and action recognition," *International Journal of Advanced Robotic Systems*, vol. 4, no. 4, pp. 301–308, 2005.

[14] F. Zacharias, C. Borst, and G. Hirzinger, "Capturing robot workspace structure: Representing robot capabilities," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 3229–3236.

A basic cognitive system for interactive continuous learning of visual concepts

Danijel Skočaj, Miroslav Janíček, Matej Kristan, Geert-Jan M. Kruijff,
Aleš Leonardis, Pierre Lison, Alen Vrečko, and Michael Zillich

***Abstract*— Interactive continuous learning is an important characteristic of a cognitive agent that is supposed to operate and evolve in an everchanging environment. In this paper we present representations and mechanisms that are necessary for continuous learning of visual concepts in dialogue with a tutor. We present an approach for modelling beliefs stemming from multiple modalities and we show how these beliefs are created by processing visual and linguistic information and how they are used for learning. We also present a system that exploits these representations and mechanisms, and demonstrate these principles in the case of learning about object colours and basic shapes in dialogue with the tutor.**

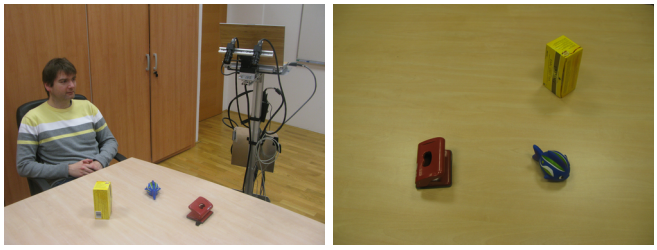
I. Introduction

An important characteristic of a cognitive system is the ability to continuously acquire new knowledge. Communication with a human tutor should significantly facilitate such incremental learning processes. In this paper we focus on representations and mechanisms that enable such interactive learning and present a system that was designed to acquire visual concepts through interaction with a human.

Such systems typically have several sources of information, vision and language being the most prominent ones. Based on the processed modal information corresponding beliefs are created that represent the robot's interpretation of the perceived environment. These beliefs rely on the particular representations of the perceived information in multiple modalities. These representations along with the cross-modal learning enable the robot to, based on interaction with the environment and people, extend its current knowledge by learning about the relationships between symbols and features that arise from the interpretation of different modalities. One modality may exploit information from another to update its current representations, or several modalities may be used together to form representations of a certain concept. We focus here on the former type of interaction between modalities and present the representations that are used for continuous learning of basic visual concepts in a dialogue with a human.

We demonstrate this approach on the robot George, which is engaged in a dialogue with the human tutor. Fig. 1 depicts a typical setup and the scene observed by the robot[1]. The main goal is to teach the robot about object properties (colours and two basic shapes). George has built-in abilities for visual processing and communication with a human, as well as learning abilities, however it does not have any model of object properties given in advance and therefore has to continuously build them. The tutor can teach the robot about object properties (e.g., 'H: This is a red thing.'), or the robot can try to learn autonomously or ask the tutor for help when necessary (e.g., 'G: Is the elongated thing red?'). Our aim is that the learning process is efficient in terms of learning progress, is not overly taxing with respect to tutor supervision and is performed in a natural, user friendly way.

In this paper we present the methodologies that enable such learning. First we present an approach for modelling beliefs stemming from multiple modalities in §II. We then show how these beliefs are used in dialogue processing in §III, followed by the description of representations and the learning process in vision in §IV. In §V we describe the system we have developed and in §VI we present an example of the scenario and the processing flow. We conclude the paper with a discussion and some concluding remarks.



(a) Scenario setup. (b) Observed scene.

Fig. 1. Continuous interactive learning of visual properties.

II. Modelling beliefs

High-level cognitive capabilities like dialogue operate on high level (i.e. abstract) representations that collect information from multiple modalities. Here we present an approach that addresses (1) how these high-level representations can be reliably generated from low-level sensory data, and (2) how information arising from different modalities can be efficiently fused into unified multi-modal structures.

The approach is based on a Bayesian framework, using insights from multi-modal information fusion [1], [2]. We

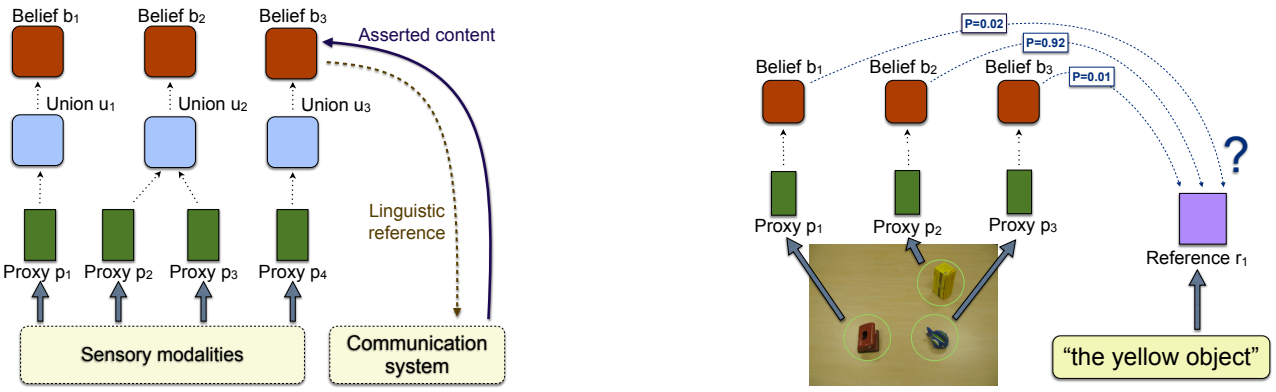
D. Skočaj, M. Kristan, A. Leonardis, and A. Vrečko are with University of Ljubljana, Slovenia, {danijel.skocaj, matej.kristan, ales.leonardis, alen.vrecko}@fri.uni-lj.si

M. Janíček, G.-J. M. Kruijff, and P. Lison are with DFKI, Saarbrücken, Germany, {miroslav.janicek, gj, plison}@dfki.de

M. Zillich is with Vienna University of Technology, Austria, zillich@acin.tuwien.ac.at

The work was supported by the EC FP7 IST project CogX-215181, and partially by the Research program Computer Vision P2-0214 (RS).

[1]The robot can be seen in action in the video accessible at http://cogx.eu/results/george.



(a) Construction of beliefs.

(b) Reference resolution for the expression "the yellow object".

Fig. 2. Multi-modal information binding: belief construction (left) and application in a reference resolution task (right).

have implemented it as a specific subsystem called the *binder* [3]. The binder is linked to all other subsystems. It serves as a central hub for gathering information about entities currently perceived in the environment. The information on the binder is inherently probabilistic, so we can deal with varying levels of noise and uncertainty.

Based on the available information, the binder seeks to fuse the perceptual inputs arising from the various subsystems, by checking whether their respective features correlate with each other. The probability of these correlations are encoded in a Bayesian network. This Bayesian network can, for example, express a high compatibility between the haptic feature "shape: cylindrical" and the visual feature "object: mug" (since most mugs are cylindrical), but a very low compatibility between "shape: cylindrical" and "object: ball".

We call the resulting (amodal) information structure a *belief*. The task of the binder is to decide which perceptual inputs belong to the same real-world entity, and should therefore be unified into a belief. The outcome of this process is a joint probability distribution over possible beliefs. These beliefs integrate the information included in the perceptual inputs in a compact representation. They can therefore be directly used by the deliberative processes for planning, reasoning and learning.

In addition to the beliefs, there are two other central data structures manipulated by the binder, proxies and unions (see also Fig. 2(a)). A *proxy* is a uni-modal representation of a given entity in the environment. Proxies are inserted onto the binder by the various subarchitectures. They are defined as a multivariate probabilistic distribution over a set of features (discrete or continuous). A *union* is multi-modal representation of an entity, constructed by merging one or more proxies. Like proxies, unions are represented as a multivariate probabilistic distribution over possible features. They are essentially a transitional layer between proxies and beliefs.

A *belief* is an amodal representation of an entity in the environment. They are typically an abstraction over unions, expressed in an amodal format. A belief encodes additional information related to the specific situation and perspective in which the belief was formed. This includes its *spatio-temporal frame* (when and where and how an observation was made), its *epistemic status* (for which agents the belief holds, or is attributed), and a *saliency value* (a real-valued measure of the prominency of the entity [4]). Beliefs are indexed via a unique identifier, which allows us to keep track of the whole development history of a particular belief. Beliefs can also be connected with each other using relational structures of arbitrary complexity.

To create beliefs, the binder decides for each pair of proxies arising from distinct subsystems, whether they should be bound into a single union, or fork into two separate unions. The decision algorithm uses a technique from probabilistic data fusion, called the *Independent Likelihood Pool* (ILP) [5]. Using the ILP, we compute the likelihood of every possible binding of proxies, and use this estimate as a basis for constructing the beliefs. The multivariate probability distribution contained in the belief is a linear function of the feature distributions included in the proxies and the correlations between these. A Bayesian network encodes all possible feature correlations as conditional dependencies. The (normalised) product of these correlations over the complete feature set provides a useful estimate of the "internal consistency" of the constructed belief.

The beliefs, being high-level symbolic representations available for the whole cognitive architecture, provide a unified model of the environment which can be efficiently used when interacting with the human user.

III. Situated Dialogue

Situated dialogue provides one means for a robot to gain more information about the environment. A robot can discuss what it sees, and understands, with a human. Or it can ask about what it is unclear about, or would like to know more about.

That makes this kind of dialogue part of a larger activity. The human and the robot are working together. They interact to instruct, and to learn more. For that, they need to build up a common ground in understanding each other and the world.

Here we briefly discuss an approach that models dialogue as a collaborative activity. It models what is being said, and

why. It enables the robot to understand why it was told something, and what it needs to do with the information.

The approach is based on previous work by Stone & Thomason [6] (S&T). In their model, an agent uses abductive inference to construct an explanation of the possible intention behind a communicative act. This intention directs how an agent's belief models need to be updated, and what needs to be paid attention to next. This kind of inference is performed both for comprehension, and for production.

The problem with S&T is that they rely on a symmetry in communication: "What I say is how you understand it." This is untenable in human-robot interaction, particularly in a setting where a robot is learning about the world. Therefore, we have adapted and extended their approach to deal with (a) the asymmetry between what has been observed fact, and what has been asserted, and (b) clarification mechanisms, to overcome breakdowns in understanding.

Algorithm 1 Continual collaborative acting

$\Sigma^\pi = \emptyset$
loop {
Perception
  $e \leftarrow$ SENSE()
  $\langle c', i, \Pi \rangle \leftarrow$ UNDERSTAND$(r, Z(c) \oplus \Sigma^\pi, e)$
  $c \leftarrow$ VERIFIABLE-UPDATE(c', i, Π)
Determination and Deliberation
  $c' \leftarrow$ ACT-TACITLY(p, c)
  $m \leftarrow$ SELECT(p, c')
  $\langle i, \Pi \rangle \leftarrow$ GENERATE$(r, c', m, Z(c) \oplus \Sigma^\pi)$
Action
  ACT-PUBLICLY$(a(i))$
  $c \leftarrow$ VERIFIABLE-UPDATE(c', i, Π)
}

Algorithm 1 presents the core of the resulting model, based on S&T. In *perception*, the agent senses an event e. It tries to understand it in terms of an intention i that results in an update of the belief model from context c to c', given the communicative resources r, possible results $Z(c)$ to use them in context c, and whatever issues are still open to be resolved Σ^π. Given the inferred intention i and potential update c' the agent then tries to carry out this update, as a *verifiable update*. To model this, we use a logical framework of multi-agent beliefs (cf. §II) that includes a notion of *assertion* [7]. An assertion is a proposition that still needs to be verified. This verification can take various forms. In George, we check whether a new piece of information can be used to consistently update a belief model (consistency), or to extend a modal model (learning) or weaken it (unlearning). Any assertion still in need of verification ends up on Σ^π.

In *deliberation*, a tacit action based on some private information p is performed by the agent. In order to make the effects c' public, a public action m is selected and performed as a realisation $a(i)$ of the generated intention to act i.

An important aspect of linking dialogue with grounded beliefs is *reference resolution*: how to connect linguistic expressions such as "this box" or "the ball on the floor" to the corresponding beliefs about entities in the environment. The binder performs reference resolution using the same core mechanisms as used for binding. A Bayesian network specifies the correlations between the linguistic constraints of the referring expressions and the belief features (particularly, the entity saliency and associated categorical knowledge). Resolution yields a probability distribution over alternative referents (see Fig. 2(b) for an example). Abductive inference then determines which resolution hypothesis to use, in the context of establishing the best explanation. This is folded together with any new information an utterance might provide, to yield an update of the robot's current beliefs.

For example, consider an utterance like "This is yellow." First, the expression "this" must be resolved to a particular, proximal entity in the environment. Resolution is performed on the basis of the saliency measures. Second, the utterance also provides new information about the entity, namely that it is yellow. The robot's beliefs get updated with this asserted information. Dialogue processing does this by selecting the belief about the referred-to entity, then incorporating the new information. Indirectly, this acts as a trigger for learning.

In George, the dynamics of assertions on Σ^π provide the main drive for how learning and dialogue interact. The vision subarchitecture can pose *clarification requests* to the dialogue system. These requests are interpreted as tacit actions (Algorithm 1), pushing an assertion onto Σ^π. This assertion may be a polar or an open statement. Then similarly to resolving any breakdown in understanding the user, the robot can decide to generate a clarification subdialogue. This dialogue continues until the (original) assertion has been verified, i.e. a proper answer has been found [8].

IV. Learning visual concepts

In the two previous sections we discussed how the modal information gathered from individual modalities is fused into unified multi-modal structures and how they are used in situated dialogue. In this section we will describe how the modal information is captured and modelled in the visual subarchitecture; how these models are initiated and how they are being continuously updated and how they can be queried to provide the abstracted information for higher-level cognitive processing.

To efficiently store and generalise the observed information, the visual concepts are represented as generative models. These generative models take the form of probability density functions (pdf) over the feature space, and are constructed in an online fashion from new observations. The continuous learning proceeds by extracting the visual data in the form of highdimensional features (e.g., multiple 1D features relating to shape, texture, colour and intensity of the observed object) and the online Kernel Density Estimator (oKDE) [9] is used to estimate the pdf in this high-dimensional feature space. The oKDE estimates the probability density functions by a mixture of Gaussians, is able to adapt using only a single data-point at a time, automatically adjusts its complexity and does not assume specific requirements on the target distribution. A particularly important feature of the oKDE is that it allows adaptation

from the positive examples (learning) as well as negative examples (unlearning) [10].

However, concepts such as *colour red* reside only within a lower dimensional subspace spanned only by features that relate to colour (and not texture or shape). Therefore, during the learning, this subspace has to be identified to provide the best performance. This is achieved by determining the optimal subspace for a set of mutually exclusive concepts (e.g., all colours, or all shapes). We assume that this corresponds to the subspace which minimises the overlap of the corresponding distributions. The overlap between the distributions is measured using the multivariate Hellinger distance [9]. An example of the learnt models is shown in Fig. 3.

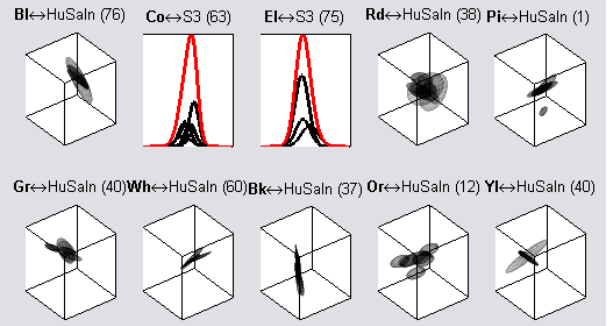


Fig. 3. Example of the models estimated using the oKDE and the feature selection algorithm. Note that some concepts are modelled by 3D distributions (e.g., "blue" which is denoted by "Bl"), while others (e.g., "compact" which is denoted by "Co") is modelled by 1D distributions.

Therefore, during online operation, a multivariate generative model is continually maintained for each of the visual concepts and for mutually exclusive sets of concepts the feature subspace is continually being determined. This feature subspace is then used to construct a Bayesian classifier for a set of mutually exclusive concepts, which can be used for recognition of individual object properties.

However, since the system is operating in an online manner, the closed-world assumption cannot be assumed; at every step the system should also take into account the probability that it has encountered a concept that has not been observed before. Therefore, when constructing the Bayesian classifier, an "unknown model" has also to be considered besides the learned models. It should account for a poor classification when none of the learnt models supports the current observation strongly enough. We assume that the probability of this event is uniformly distributed over the feature space. The a priori probability of the "unknown model" is assumed to be non-stationary and decreases with increasing numbers of observations; the more training samples the system observes, the smaller is the probability that it will encounter something novel.

Having built such a knowledge model and Bayesian classifier, the recognition is done by inspecting a posteriori probability (AP) of individual concepts and unknown model; in fact the AP distribution over the individual concepts is packed in a vision proxy, which is sent to the binder and serves as a basis for forming a belief about the observed object as described in §II (see also Fig. 2(b)).

Furthermore, such a knowledge model is also appropriate for detecting incompleteness in knowledge. It can be discovered through inspection of the AP distribution. In particular, we can distinguish two general cases. (1) In the first case the observation can be best explained by the unknown model, which indicates a gap in the knowledge; the observation should most probably be modeled with a model that has not yet been learned. A clarification request is issued that results in an open question (e.g., 'Which colour is this?'). (2) In the second case the AP of the model that best explains the observation is low, which indicates that the classification is very uncertain and that the current model cannot provide a reliable result. A clarification request is issued that results in a polar question (e.g., 'Is this red?'). In both cases, after the tutor provides the answer, the system gets the additional information, which allows it to improve the model by learning or unlearning.

V. System Architecture

We have implemented the representations and mechanisms described in the previous sections in the robot George. In this section we describe the system architecture and the individual components that are involved.

For implementation of the robot we employ a specific architecture schema, which we call CAS (CoSy Architecture Schema) [11]. The schema is essentially a distributed working memory model, where representations are linked within and across the working memories, and are updated asynchronously and in parallel. The system is therefore composed of several subarchitectures implementing different functionalities and communicating through their working memories. The George system is composed of three such subarchitectures: the *Binder SA*, the *Communications SA* and the *Visual SA*, as depicted in Fig. 4. Here, the components of the visual subsystem could be further divided into three distinct layers: the quantitative layer, the qualitative layer and the mediative layer.

In the previous subsections we assumed that the modal information is adequately captured and processed. Here we briefly describe how the relevant visual information is detected, extracted and converted in the form that is suitable for processing in the higher level processes. This is the task of *the quantitative layer* in the Visual SA. The quantitative layer processes the visual scene as a whole and implements one or more *bottom-up* visual attention mechanisms. A bottom-up attention mechanism tries to identify regions in the scene that might be interesting for further visual processing. George currently has one such mechanism, which uses the stereo 3D point cloud provided by *stereo reconstruction component* to extract the dominant planes and the things sticking out from those planes. Those sticking-out parts form spherical 3D spaces of interest (SOIs). The *SOI Analyzer* component validates the SOIs and, if deemed interesting (considering SOI persistence, stability, size, etc.), upgrades them to *proto-objects* adding information that is needed for the qualitative

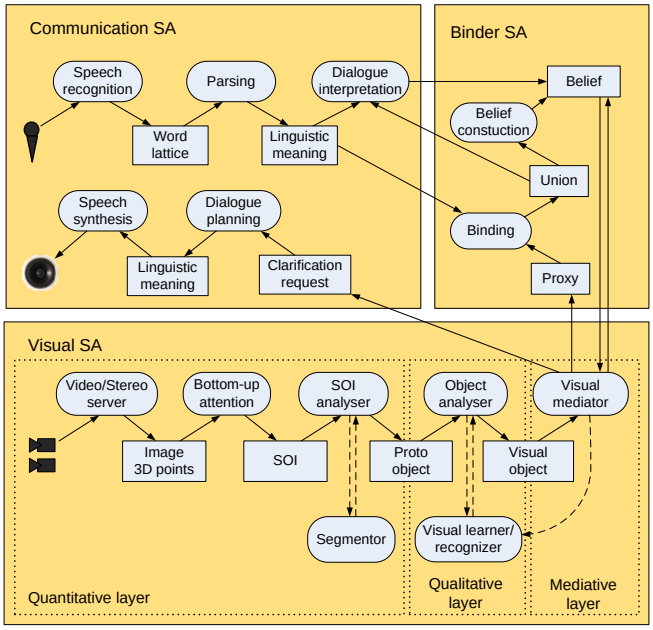


Fig. 4. Architecture of the George system.

processing, e. g. the object segmentation mask (the proto-object is segmented by the Graph cut algorithm [12] using the 3D and colour information provided by the stereo reconstruction).

The qualitative layer implements the main functionalities for recognition and learning of visual concepts that were described in §IV. This layer processes each interesting scene part (object) individually, focusing on qualitative properties. After the extraction of the visual attributes (in the *Visual Learner-recognizer*), like color and shape, the *Object Analyzer* upgrades the proto-objects to *visual objects*. Visual objects encapsulate all the information available within the Visual SA and are the final modal representations of the perceived entities in the scene. Also, the learning of visual attributes is performed in this layer.

The main purpose of *the mediative layer* is to exchange information about the perceived entities with other modalities. This is not done directly, but via the specialised a-modal subarchitecture Binder SA, that actually creates and maintains beliefs as described in §II. The *Visual Mediator component* adapts and forwards the modal information about objects to the binder (each visual object is represented by a dedicated proxy in the binder). The component also monitors beliefs for possible learning opportunities, which result in modal learning actions. Another important functionality of the mediator is to formulate and forward clarification motivations in the case of missing or ambiguous modal information. Currently, these motivations are directly intercepted by the Communication SA.

Given a clarification request, the *Communication SA* formulates a dialogue goal given the information the system needs to know and how that can be related to the current dialogue and belief-context. Dialogue planning turns this goal into a meaning representation that expresses the request in context. This is then subsequently synthesised, typically as a question about a certain object property. When it comes to understanding, the Communication SA analyses an incoming audio signal and creates a set of possible word sequences for it. This is represented as a word lattice, with probabilities indicating the likelihood that a certain word was heard, in a particular sequence. The word lattice is then subsequently parsed, and from the space of possible linguistic meaning representations for the utterance, the contextually most appropriate one is chosen [13]. Finally, dialogue interpretation takes the selected linguistic meaning. This meaning is then interpreted against a belief model, to understand the intention behind the utterance. We model this is an operation on how the system's belief model is intended to be updated with the information provided. In §VI below we provide more detail, given an example.

VI. Example scenario

A. Scenario setup

The robot operates in a table-top scenario, which involves a robot and a human tutor (see Fig. 1(a)). The robot is asked to recognise and describe the objects in the scene (in terms of their properties like colour and shape). The scene contains a single object or several objects, with limited occlusion. The human positions new objects on the table and removes the objects from the table while being involved in a dialogue with the robot. In the beginning the robot does not have any representation of object properties, therefore it fails to recognise the objects and has to learn. To begin with, the tutor guides the learning process and teaches the robot about the objects. After a while, the robot takes the initiative and tries to detect its own ignorance and to learn autonomously, or asks the tutor for assistance when necessary. The tutor can supervise the learning process and correct the robot when necessary; the robot is able to correct erroneously learned representations. The robot establishes transparency and verbalises its knowledge and knowledge gaps. In a dialogue with the tutor, the robot keeps extending and improving the knowledge. The tutor can also ask questions about the scene, and the robot is able to answer (and keeps giving better and better answers). At the end, the representations are rich enough for the robot to accomplish the task, that is, to correctly describe the initial scene.

B. Example script

Two main types of learning are present in the George scenario, which differ on where the motivation for a learning update comes from. In tutor driven learning the learning process is initiated by the human teacher, while in tutor assisted learning, the learning step is triggered by the robot.

Tutor driven learning is suitable during the initial stages, when the robot has to be given information, which is used to reliably initiate (and extend) visual concepts. Consider a scene with a single object present:

H: Do you know what this is?
G: No.
H: This is a red object.

G: Let me see. OK.

Since in the beginning, George doesn't have any representation of visual concepts, he can't answer the question. After he gets the information, he can first initiate and later sequentially update the corresponding information.

After a number of such learning steps, the acquired models become more reliable and can be used to reference the objects. Therefore, there can be several objects in the scene, as in Fig. 1, and George can talk about them:

H: What colour is the elongated object?

G: It is yellow.

When the models are reliable enough, George can take the initiative and try to learn without being told to. In this curiosity-driven learning George can pose a question to the tutor, when he is able to detect the object in the scene, but he is not certain about his recognition. As described in §IV in such *tutor-assisted* learning there are two general cases of detection of uncertainty and knowledge gaps. If the robot cannot associate the detected object with any of the previously learned models, it considers this as a gap in its knowledge and asks the tutor to provide information:

R: Which colour is this object?

H: It is yellow.

R. OK.

The robot is now able to initialise the model for yellow and, after the robot observes a few additional yellow objects, which make the model of yellow reliable enough, it will be able to recognise the yellow colour.

In the second case, the robot is able to associate the object with a particular model, however the recognition is not very reliable. Therefore, the robot asks the tutor for clarification:

R: Is this red?

H: No. This is orange.

R. OK.

After the robot receives the answer from the tutor, it corrects (unlearns) the representation of the concept of red and updates the representation of orange and makes these two representations more reliable.

In such mixed initiative dialogue, George continuously improves the representations and learns reliable models of basic visual concepts. After a while George can successfully recognise the acquired concepts and provide reliable answers:

H: Do you know what this is?

G: It is a blue object.

H: What shape is the green object?

G: It is elongated.

C. Processing flow

Here we describe the processing flow for one illustrative example. We describe in more detail what happens after the human places several objects in the scene (see Fig. 1) and refers to the only elongated object in the scene (the yellow tea box) by asserting "*H: The elongated object is yellow.*".

In the Visual SA the tea box is represented by a *SOI* on the quantitative layer, a *proto-object* on the qualitative layer and a *visual object* on the mediative layer. Let us assume that the *Visual Learner-recognizer* has recognised the object as being of elongated shape, but has completely failed to recognise the colour. In the binder this results in a one-proxy union with the binding features giving the highest probability to the elongated shape, while the colour is considered to be unknown. This union is referenced by the single robot's private belief in the belief model (Fig. 5, step 1).

The tutor's utterance 'The elongated object is yellow' is processed by the Communication SA. Speech recognition turns the audio signal into a set of possible sequences of words, represented as a word lattice. The Communication SA parses this word lattice incrementally, constructing a representation of the utterance's most likely linguistic meaning in context [13]. We represent this meaning as a logical form, an ontologically richly sorted relational structure. Given this structure, the Communication SA establishes which meaningful parts might be referring to objects in the visual context. For each such part, the binder then computes possible matches with unions present in the binding memory, using phantom proxies (Fig. 5, step 2). These matches form a set of reference hypotheses. The actual reference resolution then takes place when we perform dialogue interpretation. In this process, we use weighted abductive inference to establish the intention behind the utterance – why something was said, and how the provided information is to be used. The proof with the lowest cost is chosen as the most likely intention. Reference resolution is done in this larger context of establishing the "best explanation." Abduction opts for that referential hypothesis which leads to the overall best proof. The resulting proof provides us with an intention, and a belief attributed to the tutor is constructed from the meaning of the utterance. In our example, this attributed belief restricts the shape to elongated, asserts the colour to be yellow and references the union that includes the visual proxy representing the yellow tea box.

In the Visual SA, the mediator intercepts the event of adding the attributed belief. The colour assertion and the absence of the colour restriction in the robot's belief is deemed as a learning opportunity (the mediator knows that both beliefs reference the same binding union, hence the same object). The mediator translates the asserted colour information to an equivalent modal colour label and compiles a learning task. The learner-recognizer uses the label and the lower level visual features of the tea box to update its yellow colour model. After the learning task is complete, the mediator verifies the attributed belief, which changes its epistemic status to shared (Fig. 5, step 3). The learning action re-triggers the recognition. If the updated yellow colour model is good enough, the colour information in the binder and belief model is updated (Fig. 5, step 4).

A similar process also takes place in tutor assisted learning when the robot initiates the learning process, based on an unreliable recognition, e.g., by asking "*R: Is this red?*". In this case, the need for assistance reflects in a robot's private belief that contains the assertion about the red colour and references the union representing the object. Based on this belief, the Communication SA synthesises the above ques-

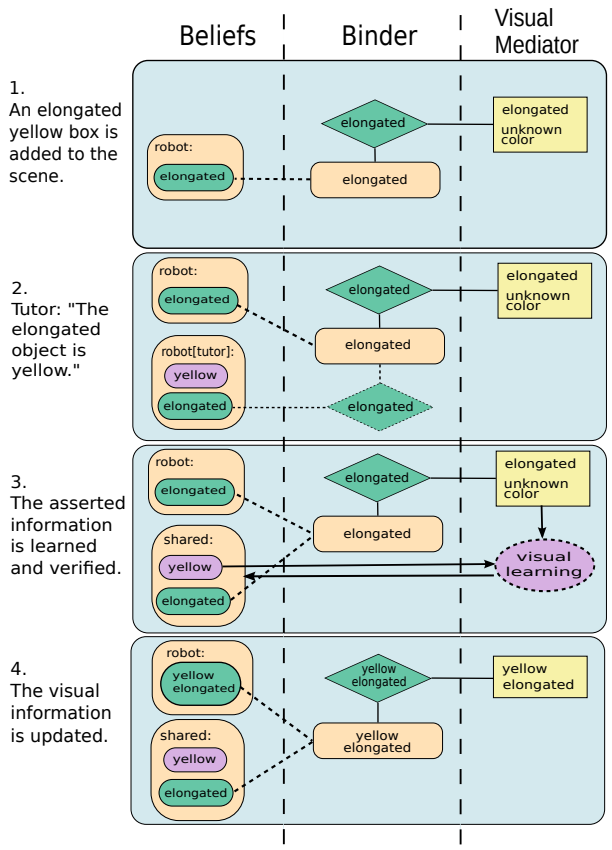


Fig. 5. Example of processing flow in the binder. The green colour represents restrictive information, while the violet colour denotes assertive information. Only the beliefs and other data structures pertaining to the yellow tea box are shown.

tion. When the robot receives a positive answer, it updates the representation of red, using a very similar mechanism as in the case of tutor driven learning.

VII. Conclusion

In this paper we presented representations and mechanisms that are necessary for continuous learning of visual concepts in dialogue with a tutor. An approach for modelling beliefs stemming from multiple modalities was presented and it was shown how these beliefs are created by processing visual and linguistic information and how they are used for learning. We also presented a system that exploits these representations and mechanisms and demonstrated these principles in the case of learning about object colours and basic shapes in a dialogue with the tutor.

We have made several contributions at the level of individual components (modelling beliefs, dialogue processing, incremental learning), as well as at the system level (by integrating the individual components in a coherent multi-modal distributed asynchronous system). Such an integrated robotic implementation enables system-wide research with all its benefits (information provided by other components), as well its problems and challenges (that do not occur in simulated or isolated environments). We are, therefore, now able to directly investigate the relations between the individual components and analyse the performance of the robot at the sub-system and system level. This will allow us to set new requirements for individual components and to adapt the components, which will result in a more advanced and robust system.

The main goal was to set up a framework that would allow the system to process, to fuse, and to use the information from different modalities in a consistent and scalable way on different levels of abstraction involving different kinds of representations. This framework has been implemented in the robot George, which is still limited in several respects; it operates in a constrained environment, the set of visual concepts that are being learned is relatively small, and the mixed initiative dialogue is not yet matured. We have been addressing these issues and the robot will gradually become more and more competent. Furthermore, we also plan to integrate other functionalities that have been under development, like motivation and manipulation.

The presented system already exhibits several properties that we would expect from a cognitive robot that is supposed to learn in interaction with a human. As such, it forms a firm basis for further development. Building on this system, our final goal is to produce an autonomous robot that will be able to efficiently learn and adapt to an everchanging world by capturing and processing cross-modal information in an interaction with the environment and other cognitive agents.

References

[1] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[2] R. Engel and N. Pfleger, "Modality fusion," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed. Berlin: Springer, 2006, pp. 223–235.

[3] H. Jacobsson, N. Hawes, G.-J. Kruijff, and J. Wyatt, "Crossmodal content binding in information-processing architectures," in *Proc. of the 3rd International Conference on Human-Robot Interaction*, 2008.

[4] J. Kelleher, "Integrating visual and linguistic salience for reference resolution," in *Proceedings of the 16th Irish conference on Artificial Intelligence and Cognitive Science (AICS-05)*, N. Creaney, Ed., 2005.

[5] E. Punskaya, "Bayesian approaches to multi-sensor data fusion," Master's thesis, Cambridge University Engineering Department, 1999.

[6] R. Thomason, M. Stone, and D. DeVault, "Enlightened update: A computational architecture for presupposition and other pragmatic phenomena," in *Presupposition Accommodation*, to appear.

[7] M. Brenner and B. Nebel, "Continual planning and acting in dynamic multiagent environments," *Journal of Autonomous Agents and Multi-agent Systems*, 2008.

[8] G. Kruijff and M. Brenner, "Phrasing questions," in *Proceedings of the AAAI 2009 Spring Symposium on Agents That Learn From Humans*, 2009.

[9] M. Kristan and A. Leonardis, "Multivariate online kernel density estimation," in *Computer Vision Winter Workshop*, 2010, pp. 77–86.

[10] M. Kristan, D. Skočaj, and A. Leonardis, "Online kernel density estimation for interactive learning," *Image and Vision Computing*, 2009.

[11] N. Hawes and J. Wyatt, "Engineering intelligent information-processing systems with CAST," *Advanced Engineering Infomatics*, vol. 24, no. 1, pp. 27–39, 2010.

[12] Y. Boykov, O. Veksler, and R. Zabih, "Efficient approximate energy minimization via graph cuts," *PAMI*, vol. 20, no. 12, pp. 1222 – 1239, 2001.

[13] P. Lison and G. Kruijff, "Efficient parsing of spoken inputs for human-robot interaction," in *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 09)*, Toyama, Japan, 2009.

Identifying and Resolving Ambiguities within Joint Movement Scenarios in HRI

Maryamossadat N. Mahani and Elin Anna Topp

Abstract— **In this work we report on enhancing interactions of users with a service robot by making the robot's interface account for its operational requirement and also by making its interface provide a more clear indication of the robot's internal perceptions to the users. The work was initiated by observation of several cases of a specific human-robot interaction problem in an exploratory user study with a "robot follows user" scenario. In this paper we present our research on analysis of the observed problem situations, the respective suggestions for improvements of the used system, and evaluation of the subsequent modifications of the system. The results show significant improvements in interaction of users with the robot.**

I. INTRODUCTION

Important to the successful application of robots by arbitrary users who are not designers of the system, is their "ease of use". Two primary factors affecting the ease of use of robots are effective communication and intuitive interaction with them. An intuitive interaction in this case refers to an interaction which can be related to interaction with other already adopted machines, or to the interaction with humans.

In an exploratory user study [8], referred to as "pilot study", in which a group of persons with different experience with robots were supposed to guide a robot around in an indoor environment, we identified a pattern of interaction problem. We termed the problem a "deadlock", since it is similar to a synchronization problem occurring with two program threads blocking each other. In such a deadlock situation the interaction flow was interrupted due to a mismatch of the mental model the users had of the robot with the actual state of the system. This happened despite the facts that the robot gave verbal feedback to the issued command and that the users had been informed about how to fulfill certain requirements to proceed. It was thus obvious that the system did not fulfill any "ease of use" requirements, since it was not at all intuitive to interact with, although it could be spoken to and gave spoken feedback.

A significant amount of research to date has been focused on enabling robots to recognize human gestural cues. Equally important is equipping robots with gestures that are understandable and intuitive to the users. A strong factor in human-human communication is spoken language, but this is usually supported and enhanced by non-verbal communication, i.e., "body language". In a study with a robot penguin, Sidner *et al.* [7] could show that this applies also to human-robot communication in the sense that users get more engaged in a conversation with a robot when this robot stresses its verbal utterances with adequate movements, e.g., nods.

In addition to the influence of body language in communication, it is important to cover the "gulf of evaluation" [6] for the users so they do not face difficulty in evaluating whether the response of a machine meets the desired goal. Humans are quite used to the fact that "pressing a button" will result in the respective device starting to "do something", otherwise they will assume that something is wrong. Hence, giving a command to a robot should result in the robot starting to perform the task related to this command or give appropriate feedback that helps the user understand why the task is not performed – as long as the system is working properly.

The need for greater intelligence of robots and command interfaces that adapt to situations is deemed necessary for successful application of robots [2]. Goodfellow *et al.* [3] show the results of case studies in situations where a robot requires further information for performing a certain task and its interface is designed to let the robot and user communicate about it. In this work, we focus on communicating an operational requirement of a robot to its users by making its interface account for the non-satisfied requirement.

Assuming that we could apply the idea of the body language and some more sophisticated verbal feedback – a "spatial prompt" [4], that would explain and resolve the situation, we designed and evaluated a system improvement to handle the deadlock situations we had identified.

II. THE RESEARCH PROBLEM

The platform of this study is a Performance PeopleBot [1]. In the context of this work, it is introduced as a service robot that is supposed to provide services in personal contexts, such as houses or office environments. Table 1. represents a list of some of the possible commands [2] to the robot and the robot's response to them. The robot interacts with the surroundings using its sensors. Fig. 1 shows the robot's sensory detection range. The robot can detect and track a user as long as the user is in front of the robot's baseline and within its navigation range. Inside the robot's control software, a parameter defines a minimal distance threshold to the user that the robot should not move beyond. This "safety distance" parameter is set to one meter in this work and is meant to avoid the robot from bumping into the users

M. N. Mahani is a PhD student at the Electrical Engineering and Computer Science department, University of Kansas, Lawrence, KS, USA mahani@eecs.ku.edu

E. A. Topp is a post-doctoral researcher at the Computer Science Department, Lund University, Lund, Sweden Elin_Anna.Topp@cs.lth.se

[1] http://www.mobilerobots.com/ResearchRobots/

[2] The complete list can be found at: http://people.ku.edu/∼mahani/commandsList.pdf

TABLE I

COMMANDS AND FEEDBACK OF THE ROBOT.

Robot's sensory detection	**User command**	**Robot's verbal feedback**	**Robot's action**
A person is detected in proximity of the robot	———	Hello, my name is Minnie, show me around please	———
Keeping track of the user	Follow me	I will follow you	If user is far enough, follows the user else stays at its current place
Keeping track of the user	Stop	Stopped following	Stops moving
Lost track of the user	——-	I think I lost you	No action
Regained track of the user	——-	I found you again	——-

or making them feel unsafe. In all the studies reported in this work, the subjects were given instructions about the robot's operation specifications, including its detection range and the safety distance.

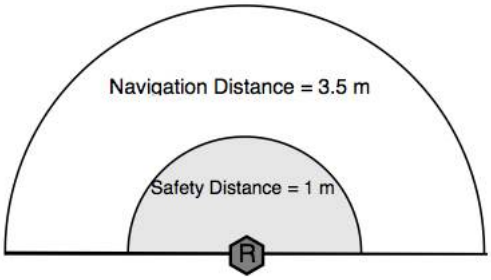


Fig. 1. The robot's sensory detection range

III. IDENTIFYING INTERACTION AMBIGUITIES

The pilot study experiments were video taped. A video analysis of this data was used to identify cases of interaction problem, based on the uneven turn takings and situations of trouble and repair that they caused.

A. The deadlock situation

A common pattern of interaction problem was identified as part of a "user-following" scenario, during which the user gives a command to the robot, the robot responds that it will perform the user's request but it does not carry out any actions towards fulfillment of the command. As a result, the user waits for a while, repeats the same command, gets the same response back, ..., until the situation is resolved by "meta-communication" between experiment leader and user. This problem happened to 4 out of the 5 subjects of the study at the initiation of a "follow me" command.

1) Under the scene: What is happening inside the robot's control software during a deadlock situation, explains the above problem. The robot has received the command, but the user is standing closer than the safety distance to the robot, the robot is waiting for the user to start moving further, and the user has forgot about the safety distance requirement.

2) Lack of clear indication of the robot's perception: Occurrence of the deadlock situation could also be affected by one of the robot's interaction specifications, that is the robot will not try to orient towards a user when having a conversation. The focus in the initial version of the robot's software was mostly on being able to autonomously follow a user, so it was only when running the system with arbitrary users that such interaction issues became apparent. Fig. 2 shows a sample relative position of a user and the robot having a conversation. The possible effect of this behavior of the robot on user's perception of the robot's status is further supported by the observed behavior of a few subjects who tried to fix the deadlock situation by going to the front of the robot and repeating the command.

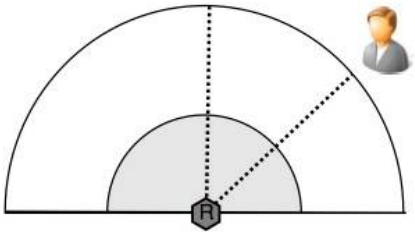


Fig. 2. Sample relative orientation of robot and user during a conversation

IV. RESOLVING INTERACTION AMBIGUITIES

A. Explicit presentation of the robot's internal processing

The only explicit feedback of the robot in response to the "follow me" command has been verbal output. At the same time, the robot does not have any means of showing the user that it is still tracking him even though it might not be directly facing the user. We propose that a more believable and intuitive behavior of the robot can be achieved by making the robot show the users that it is attending them not only verbally but also physically.

As a solution to the deadlock situation, we equipped the robot with a turning motion that would orient the robot towards the user until a maximum angular difference of 6 degrees is reached. We assumed that this will convey the robot's attention to the users and as a result it will make the robot's behavior more intuitive. The orientation motion is performed before uttering the "I will follow you" feedback.

B. Robot's Interface accounting for its operational requirements

When a deadlock situation happens, the robot does not provide any feedback to the user about its non-satisfied requirement. It should be noted again that the subjects were given instructions about the safety distance requirement before start of trials. We propose that making the robot's interface stand for its non-satisfied requirement will assist in keeping a smooth flow of interaction and help the users take a corrective action. As a solution, the robot is made to provide a verbal "spatial prompt" about its non-satisfied requirement

in a deadlock situation. An instructive verbal feedback is used which includes directives about the corrective action that should be taken and the reason behind it. The verbal feedback is: "You are too close to me, you have to move a bit further and I will follow you."

V. EVALUATION AND RESULTS

A laboratory study with several users was conducted to evaluate the solutions for the deadlock situation. The results from the laboratory study were used in forming a strategy based on small robot orientation movements and "spatial prompting". The solution strategy was further applied to a prototypical user study tool and tested with users in domestic contexts.

A. Laboratory user study

One of our main evaluation goals was getting to know the subjects' perception of the robot's behavior, that is what they think the robot is doing or is trying to do. "A much more reliable and possibly objective method for measuring the users' perception and cognition is to observe their behavior" [1]. At the same time to reach our evaluation goals, it was necessary to employ a qualitative evaluation method rather than any quantitative cross-validation. A total of 20 cooperative user trials were performed. Since we were interested in identifying occurrence of specific, known error situations, the observations were noted using pen and paper. Each trial run was followed by a combination of a post-task walkthrough and interview to reflect actions of the subjects to them and gain a correct interpretation of their behavior.

1) Results of the laboratory study: The deadlock situation happened in 4 out of 20 studies, but in all cases the users took the right corrective action to fix the situation upon hearing the verbal feedback. The subjects described their interpretation of the orientation motion either as a sign of attention of the robot or its state of being ready to move.

B. Study in users' homes

The laboratory study results proved that the improvements of the "user-following" behavior enhanced the interactions. The suggested improvements were also implemented in an extended study tool version of the framework for Human Augmented Mapping (interactive robotic mapping), which was used for a further, exploratory study in users' homes [5], with 8 trials in 7 homes. In particular, the robot was turning towards the assumed interaction partner before stating that it would follow. Also the utterance "Please move on, you are too close to me!" was given when the situation required. The utterance was in this implementation controlled by the experiment leader which allowed to use the spatial prompt not only when the "follow-me" command had been issued, but also in any cases of "arising deadlocks", that is whenever user and robot were about to get stuck in their joint movement. The experiment leader decided in most cases based on the distance of the user to the robot, i.e., when the user was in fact too close to the robot for any movement considering the assumed safety distance. Another type of case was when a tracking failure was obviously repairable by having the user move a little to be visible as "moving person" again.

1) Results of the home study: From the observations made in the study in users' homes it is obvious that those can only be seen as anecdotic evidence of the success of our subtile changes to the "user-following" behavior. However, we can state a clear improvement in comparison to the initial pilot study, in the sense that *no actual deadlocks* having to be resolved by "meta-communication" between experiment leader and user were observed. Further, the utterance "Please move on, you are too close to me!" caused the respective subject in 9 cases of arising deadlocks to move further away from the robot. In most of the cases the problem was solved with the first prompt, i.e., the robot was able to move afterwards, and after two prompts at most the intended joint movement was possible in all cases.

VI. CONCLUSIONS

The results of both user studies show a considerable decrease in occurrences of deadlocks. These results indicate that the orientation motion improved the interactions by making the robot's behavior more intuitive. Also, the "spatial prompt" helped in keeping a smooth interaction flow by making the users take a corrective action when necessary. Overall, these results put emphasis on the need for the robot to communicate its internal processing clearly to the users so that the users keep a correct mental model of the robot's status. The results also emphasize the positive effects of the robot's interface accounting for its operational requirements in enhancing human robot interactions.

REFERENCES

[1] C. Bartneck, E. Croft, D. Kulic "Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots". *In Proceedings of the Metrics for Human-Robot Interaction Workshop in affiliation with the 3rd ACM/IEEE International Conference on Human-Robot Interaction*, University of Hertfordshire, Amsterdam, 2008, pp. 37-44.

[2] K. Gold, "An information pipeline model of human-robot interaction", *In Proceedings of the 4th ACM/IEEE conference on Human-Robot Interaction*, La Jolla, CA, USA, 2009, pp. 85-92.

[3] I. J. Goodfellow, N. Koenig, M. Muja, C. Pantofaru, A. Sorokin, L. Takayama, R. Maier, "Help me help you: interfaces for personal robots", *In proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, Osaka, Japan, 2010, pp. 187-188.

[4] A. Green, H. Hüttenrauch, "Making a Case for Spatial Prompting in Human-Robot Communication", *Multimodal Corpora: From Multimodal Behaviour theories to usable models, workshop at the Fifth international conference on Language Resources and Evaluation*, Genova, 2006.

[5] H. Hüttenrauch, E. A. Topp, K. Severinson Eklundh, "The Art of Gate-Crashing: Bringing HRI into Users' Homes", *In Robots in the Wild: Exploring Human-Robot Interaction in Naturalistic Environments, Journal of Interaction Studies* 10(3), 2009, pp. 274-297.

[6] Norman D A (1998) The Design of Everyday Things. New York, Doubleday

[7] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, C. Rich, "Explorations in Engagement for Humans and Robots", *Artificial Intelligence*, 2005.

[8] E. A. Topp, H. Hüttenrauch, H. I. Christensen and K. Severinson Eklundh, "Bringing Together Human and Robotic Environment Representations - A Pilot Study", *In Proceedings of the IEEE/RSJ Conference on Intelligent Robots and Systems*, Beijing, China, 2006, pp. 4946-4952.

The "Curious Robot" learns grasping in multi-modal interaction

Ingo Lütkebohle, Julia Peltason, Robert Haschke, Britta Wrede and Sven Wachsmuth

I. Introduction

Practical robots have come a long way, from being confined to strong cages in industrial manufacturing halls to open environments shared with humans. One consequence if robots are to share spaces with humans is that they must be able to learn from them – that much is well accepted. The reverse – that the robot becomes the "teacher" and the human the "student" – is less commonly seen, however. This is despite the fact that many applications tacitly assume that humans learns about the robot, e.g. from a manual or through instruction by an expert.

We surmise that there is a great deal of potential in an *explicit* reversal of the roles. Therefore, we have investigated how this reversal of the traditional roles can improve HRI. Concretely: *How could a robot structure the dialog such that a naive human partner is aware of her/his possible dialog actions?* The goal is to make humans able to act with confidence despite having absolutely no prior knowledge of either the robot's goals or its capabilities.

To achieve this ambitious goal, several hard problems must be adressed. One important issue is the vocabulary problem [1], that describes the fact that humans do not know what the system understands, in particular at the beginning of an interaction [2]. Another well known problem is that user's expectations about a system are strongly shaped by appearance [3], [4], which may lead to erroneous assumptions [5]. Last, but not least, it is not clear how to provide guidance in an easy to understand way and this requires an iterative, study-based approach towards system development [6].

To investigate how robot guidance can improve upon this, we have introduced the "Curious Robot" interactive scenario for learning about real-world objects [5]. In it, we have used a mixed-initiative [7] approach, that has the robot query the human for information at appropriate points during the interaction. For example, the robot queries a human about object labels and how to grasp an object. Initiative is guided by visual saliency information [8].

In this scenario, our results indicate that closed questions provide excellent guidance to the human, resulting in confident and very consistent (across subjects) answers. The reverse has also been found, with partially open questions leading to considerable more confusion and inconsistency.

This work was partially supported by the German Federal Ministry of Education and Research (BMBF) under grant no. 01IME01N and the German Research Society (DFG)within the Collaborative Research Centre 673 "Alignment in Communication".

Ingo Lütkebohle, Julia Peltason and Britta Wrede are with the Applied Informatics Group, Bielefeld University, Germany. {iluetkeb,jpeltaso,bwrede}@techfak.uni-bielefeld.de

Robert Haschke is with the Neuroinformatics Group, Bielefeld University, German. {rhaschke}@techfak.uni-bielefeld.de

Sven Wachsmuth is with the Central Lab Facilities, Cognitive Interaction Technology Excellence Cluster, Bielefeld University, Germany. {swachsmu}@techfak.uni-bielefeld.de

II. Multi-modal interaction

A particular problem during speech based interaction is that many activities are hard to describe verbally. Instead, we found that participants prefer a mixture of demonstration and description [5]. Therefore, we have now added hand-posture sensing as an input device to describe grasping using a CyberGlove.

An issue with posture sensing through a glove is to determine when to use it, particularly when motions are only mimicked for demonstration. In the video, we demonstrate how verbal and haptic information are combined to overcome this issue.

III. Progress inquiries

Not surprisingly, we found subjects in the learning scenario to be interested in knowing what the robot has learned. This led them they to interrupt the current activity through questions about the system's knowledge and the current state.

We can accomodate this through our grounding-based dialog [9], which allows nesting of individual exchanges and demonstrate this capability in the video. To encourage subjects to ask questions, the system has also been equipped with voice activity detection, to slow down upon sensing speech. This provides users with feedback that their question is currently possible and the system is attending. The latter aspect is also supported through gaze feedback.

IV. Conclusion

We describe how to extend a scenario based on the idea of robot guidance with posture sensing for multi-modal descriptions and improved learning feedback. We also summarize user studies on a previous iteration of the scenario to motivate the chosen approach.

The video is available at `http://aiweb.techfak.uni-bielefeld.de/cr-icair-2010`.

References

[1] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, November 1987. [Online]. Available: http://dx.doi.org/10.1145/32206.32212

[2] M. Hanheide and G. Sagerer, "Active memory-based interaction strategies for learning-enabling behaviors," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Munich, 2008.

[3] M. Lohse, F. Hegel, and B. Wrede, "Domestic applications for social robots - a user study on appearance and function," *Journal of Physical Agents*, vol. 2, pp. 21–32, 2008.

[4] M. Lohse, "The role of expectations in hri," in *New Frontiers in Human-Robot Interaction*, 2009.

[5] I. Lütkebohle, J. Peltason, L. Schillingmann, C. Elbrechter, B. Wrede, S. Wachsmuth, and R. Haschke, "The Curious Robot - Structuring Interactive Robot Learning," in *International Conference on Robotics and Automation*, Robotics and Automation Society. Kobe, Japan: IEEE, May 2009.

[6] M. Lohse, M. Hanheide, K. Rohlfing, and G. Sagerer, "Systemic interaction analysis (sina) in hri," in *Conference on Human-Robot Interaction (HRI)*, IEEE. San Diego, CA, USA: IEEE, 11/03/2009 2009, pp. 93–100.

[7] J. F. Allen, "Mixed-initiative interaction," *IEEE Intelligent Systems*, vol. 14, no. 5, pp. 14–23, 1999.

[8] Y. Nagai, K. Hosada, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, December 2003. [Online]. Available: http://dx.doi.org/10.1080/09540090310001655101

[9] J. Peltason, F. H. Siepmann, T. P. Spexard, B. Wrede, M. Hanheide, and E. A. Topp, "Mixed-initiative in human augmented mapping," in *International Conference on Robotics and Automation*, IEEE. Kobe, Japan: IEEE, 14/05/2009 2009.