The Vernissage Corpus: A Conversational Human-Robot-Interaction Dataset

Dinesh Babu Jayagopi[1], Samira Sheiki[1,2], David Klotz[3], Johannes Wienke[3], Jean-Marc Odobez[1,2], Sebastien Wrede[3], Vasil Khalidov[1], Laurent Nyugen[1,2], Britta Wrede[3], Daniel Gatica-Perez[1,2]
[1]Idiap Research Institute [2]École Polytechnique de Fédérale de Lausanne (EPFL), Switzerland
[3]Bielefeld University, Germany.
{djaya, samira.sheiki, odobez, vasil.khalidov, lnguyen, gatica}@idiap.ch
{dklotz@cor-lab, jwienke@techfak, sebastian.wrede, bwrede@techfak}.uni-bielefeld.de

***Abstract*—We introduce a new conversational Human-Robot-Interaction (HRI) dataset with a real-behaving robot inducing interactive behavior with and between humans. Our scenario involves a humanoid robot NAO[1] explaining paintings in a room and then quizzing the participants, who are naive users. As perceiving nonverbal cues, apart from the spoken words, plays a major role in social interactions and socially-interactive robots, we have extensively annotated the dataset. It has been recorded and annotated to benchmark many relevant perceptual tasks, towards enabling a robot to converse with multiple humans, such as speaker localization and speech segmentation; tracking, pose estimation, nodding, visual focus of attention estimation in visual domain; and an audio-visual task such as addressee detection. NAO system states are also available. As compared to recordings done with a static camera, this corpus involves the head-movement of a humanoid robot (due to gaze change, nodding), posing challenges to visual processing. Also, the significant background noise present in a real HRI setting makes auditory tasks challenging.**

Keywords*—HRI corpus; Multimodal dataset; Social-robotics; Vernissage.

I. Introduction

One of the fundamental challenges in HRI is providing humanoid robots with the audio-visual perception capabilities to interact with multiple human partners [3]. Towards this goal, realistic interaction scenarios need to be studied, wherein the nonverbal behavior of the humanoid robot induces nonverbal behavior of the humans, e.g. looking towards a picture when the robot indicates this or looking at the human partner when discussing a painting. To enable perception in such realistic scenarios, new methods in processing unimodal and multimodal data need to be developed. Furthermore, existing methods have to be adapted and redesigned to meet the challenges that accompany recording with the audio-visual sensors on a humanoid robot. In order to be a realistic interaction partner, a humanoid robot needs to perform appropriate actions, for example nodding, or gaze changes which affect the sensing process. Though gaze change is desirable from an interaction perspective such behavior severly degrades the sensing quality, as the sensor is moved and motor noise is added. Also, the sensing, computing, and communication capabilities on the robot are limited and constrain each other.

Although humans rely heavily on the information provided in the verbal channel, robotics perceptual research, still mainly focusses on visual aspects such as object recognition or navigation (e.g. [9]). Among those corpora that have focused on audio-visual perception tasks [6], [2], [1] in a conversational scenario, none of them have all the advantages of our dataset: an interesting scenario, more than one interaction partner, a commercially available robot (with consumer sensors rather than high-end sensors), extensive annotations. NAO's commercial availability facilitates repeatability and comparison of experiments. We believe that the scenario, where NAO acts as an art guide and a quiz master, is both an interesting and reasonably controlled application scenario of humanoid robots. While datasets exist for benchmarking a single perceptual task, our dataset allows benchmarking several perceptual tasks on a single dataset in a more realistic situation.

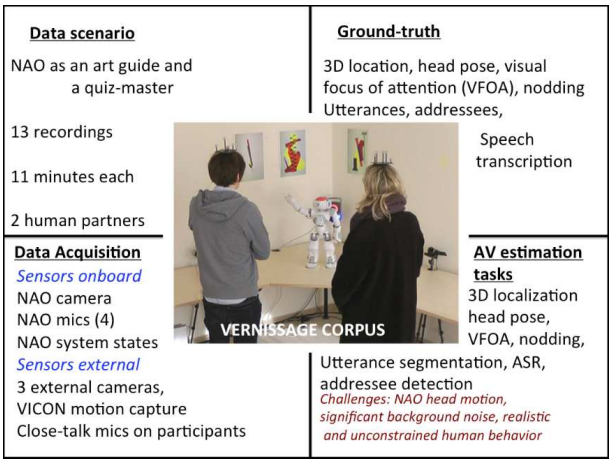

Figure 1. Overview of the Vernissage corpus: scenario, various modalities, annotations, and possible audio-visual perception tasks.

II. Dataset

Scenario: Our dataset (called 'The Vernissage[2] corpus') has 13 sessions of NAO interacting with two persons. The robot serves as an art guide, explaining the paintings to the participants and then quizzing them in art and culture. The scenario involves a stationary robot, exhibiting significant gesticulations to facilitate the interaction, for example, turning its head, nodding, and pointing its hand towards objects of interest. A Wizard-of-Oz was used to manage the dialog as well as the robot's gaze and nodding. The behavior of the

[1]http://www.aldebaran-robotics.com

[2]*vernissage*: French for the opening of an art exhibition

human partners was unconstrained. Each interaction lasted around 11 minutes. Fig. 1 gives an overview of the corpus.

Our scenario is an instance of the Vernissage scenario, inspired by a recent work that has studied and documented human interaction experiences with NAO as an art guide in a German art museum [7]. The Vernissage scenario was chosen as it offers sufficient flexibility as well as control over the human-robot interaction. It allows for a continuous change in difficulty w.r.t robot/mixed initiative, mild to intensive user involvement, number of participants, and the text uttered by the robot. Furthermore, as the robot is stationary, the complexity involved in adapting and extending existing perception methods is reasonable, but still challenging.

Acquisition: The dataset comprises a synchronized multimodal corpus with multiple auditory, visual, and robotic system information channels. The recording method is discussed elsewhere [8], and is inspired by the SInA method proposed in [5] which focuses on synchronizing internal logging data with external manually annotated data in order to analyze specific issues of HRI. NAO video data is mono at VGA resolution and audio data comes from four microphones. To have ground-truth information for all the audio-visual processing tasks, three close-field external cameras, a motion capturing system and close-talk microphones on the participants were deployed.

Annotation: The corpus is annotated with several nonverbal cues such as speech utterances, 2D head-location, nodding, visual focus of attention (VFOA), and addressees. Speech transcription is also available. The reliabilities of slightly subjective annotations such as addressees, visual focus of attention, and nodding were studied on a subset of the dataset using a secondary annotator. The reliabilities were sufficiently high, thereby facilitating the possibility of training models using these annotations as ground-truth. Further statistics about the annotations is available with the dataset. A research report with full description of the dataset and the annotation statistics is also available [4]. Apart from the richness of the sensor data and the annotations on that data, the dataset includes the robot system data consisting of the robot's 3D location of its body, joint angles, wizard commands, internal events for speech and gesture production, usage of CPU, memory, and battery. The motion capturing system gives the 3D location of the participants and their head-pose.

III. Dataset Use-cases

Our corpus allows benchmarking of several perceptual tasks in a realistic HRI context. For example, speaker localization and speech segmentation using NAO microphones; head tracking, pose estimation and VFOA estimation using the monocular video. Methods to detect the target of the speech utterances i.e. addressees could also be evaluated.

Some of the perceptual tasks could make use of inputs from different sources in single modalities for comparison. For example, audio tasks such as utterance estimation could use NAO's microphones or close-talk microphones. When performing automatic speaker localization using NAO's microphone data, we can benchmark the performance of the processing methods using the utterance annotation. The loss of performance when using the close-talk microphones and different NAO microphones tells us how challenging the task is. Studying this, better speech enhancement techniques could be devised to improve the signal-to-noise ratio. Visual tasks such as head-pose estimation and head-tracking could use video from NAO's camera and compare with other external cameras. In certain cases, the effect of errors in estimating a low-level cue such as speaker localization on estimating a high-level cue such as addressees could be studied.

Audio-visual tasks can also be attempted, e.g., audio-visual people tracking, that could exploit both audio data (i.e. speaking information) and video data (i.e. head motion information). Addressee estimation requires both utterance information as well as gaze information. Both nodding estimation and prediction can make use of audio-visual cues such as speaking cues of others and head-movement of self.

Apart from the audio-visual modalities and NAO system data, other automatically estimated cues or their ground-truth could also serve as an important context for some tasks such as VFOA and addressee estimation. For example, VFOA estimation could be improved with the dialog state of NAO (e.g., which painting he is talking about). Wizard commands and their timing could be used to study how to automate gaze-shift and nodding. For estimating a higher-level cue such as 'who is being addressed', lower-level cues such as VFOA, and dialog context (from both NAO as well as the participants) could be relevant. It would be interesting to study how useful each of these cues is (using the ground-truth), what the degradation when using automatically extracted cues is, and also study what is the best way of fusing the information.

IV. Conclusion

In this paper, we presented a new HRI corpus towards analyzing multi-party interaction behavior and enabling robust multimodal perception methods. The perception methods need to work well in challenging sensing conditions, that accompany a realistic yet challenging scenario, where the humanoid robot exhibits significant nonverbal behaviors, and recording with a commerical-quality robot sensors. The dataset will be available in March 2013.

Acknowledgment: This research was funded by the EU HUMAVIPS project.

References

[1] X. Alameda-Pineda et al. The RAVEL data set. In *ICMI 2011 Workshop on Multimodal Corpora*, Alicante, Spain, Nov 2011.

[2] E. Arnaud et al. The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements. In *Proc. ICMI*. ACM, 2008.

[3] T. Fong et al. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3):143–166, 2003.

[4] D. Jayagopi et al. The vernissage corpus: A multimodal human-robot-interaction dataset. In *Idiap research report (Idiap-RR-33-2012)*, 2012.

[5] M. Lohse et al. Systemic interaction analysis (SInA) in HRI. In *Proc. Human-Robot Interaction (HRI)*, San Diego, CA, USA, 2009.

[6] Y. Mohammad et al. The h3r explanation corpus human-human and base human-robot interaction dataset. In *Proc. ISSNIP*. IEEE, 2008.

[7] K. Pitsch et al. Attitude of german museum visitors towards an interactive art guide robot. In *Proc. HRI*. ACM, 2011.

[8] J. Wienke et al. A framework for the acquisition of multimodal HRI data sets with a whole-system perspective. In *LREC*, 2012.

[9] Z. Zivkovic et al. From sensors to human spatial concepts. *Robotics and Autonomous Systems*, 55(5):357–358, 2007.