FAIM/ISCA Workshop on Artificial Intelligence for Multimodal Human Robot Interaction
14-15 July 2018, Stockholm, Sweden

Demonstrating and Learning Multimodal Socio-communicative Behaviors for HRI: Building Interactive Models from Immersive Teleoperation Data*

Gérard Bailly & Frédéric Elisei
GIPSA-lab, Univ. Grenoble-Alpes, CNRS, Grenoble INP, Grenoble - France
firstname.lastname@gipsa-lab.fr

Abstract

The main aim of artificial intelligence (AI) is to provide machines with intelligence. Machine learning is now widely used to extract such intelligence from data. Collecting and modeling multimodal interactive data is thus a major issue for fostering AI for HRI. We first discuss the egg-and-chicken problem of collecting ground-truth HRI data without actually disposing of robots with mature social skills. Particular issues raised by the current multimodal end-to-end mapping frameworks are also commented. We then analyze the benefits and challenges raised by using immersive teleoperation for endowing humanoid robots with such skills. We finally argue for establishing stronger gateways between HRI and Augmented/Virtual Reality research domains.

1 Introduction

Endowing humanoid robots with appropriate multimodal socio-communicative and task-specific behaviors for convincing Human-Robot interaction (HRI) is a challenging issue. The classical approach consists in scaling behaviors collected during Human-Human interaction (HHI). This scheme faces two important issues: (a) the impoverished or augmented sensorimotor abilities of robots that require to map between different scores and reconsider spatio-temporal patterns; (b) the drastic change of observed human behaviors in front of avatars as compared with humans. Adapting HHI models to HRI condition is not straightforward since social rewards are also difficult to objectify.

This paper has four main sections. We first discuss the problem of collecting relevant HRI data in section 2. We then sketch in section 3 the benefits and challenges of using immersive teleoperation for teaching multimodal socio-communicative behaviors to humanoid robots. Section 4 introduces the Machine Learning (ML) techniques used to build behavioral models for autonomous HRI using these collected data. We finally discuss challenging issues raised by this data-driven framework.

*This work is supported by ROBOTEX (ANR-10-EQPX-44-01), SOMBRERO (ANR-14-CE27-0014), PERSYVAL (ANR-11-LABX-0025) and by the TENSIVE project.

2 Collecting interactive multimodal data

Several pathways have been explored so far to collect relevant interaction data, as needed by the HRI model-building algorithms.

2.1 Interactive human-human data

Most interactive multimodal HRI behavioral models are built from rules picked up in the literature – found in bibles such as Kendon [2004] for gesturing or Kita [2003] for pointing – or multimodal data collected during dyadic HHI [Bilakhia *et al.*, 2015] or group interviews [Oertel *et al.*, 2014].

The collected HHI signals of the source participant are then either directly scaled to perceptuo-motor abilities of the target robot or first converted to stamped perceptuo-motor events (such as "say X to Y at time T", "look at Y at time T", etc; see the SAIBA framework below) which then trigger robot-specific motor primitives/programs (eg. repositories of gestures [Krenn and Pirker, 2004] or gesture controllers [Nguyen *et al.*, 2017b]). This re-targeting of HHI data to virtual agents is quite straightforward. However, this operation is much more difficult for robots, whose kinematics and dynamical behavior strongly impact the control strategy.

2.2 HRI multimodal data from external views

Numerous datasets have been collected to observe human behaviors when conversing with autonomous virtual agents [McKeown *et al.*, 2012] or robots [Castellano *et al.*, 2010]. While these datasets are very informative about users' expectations and deceptions, they do not actually provide data scientists with signals that can be directly exploited by endogenous sensorimotor capabilities of robots.

2.3 Multimodal data from the robot's point of view

Few datasets have been collected from the robot's point of view (POV) since this presupposes that the experiments at least involve a robot that passively experience"s the interaction. As an example, Azagra et al [Azagra *et al.*, 2017] dataset contains recordings of several users teaching different object classes to the robot Baxter. But the robot remains inactive : it passively experiences the interaction.

2.4 Perception in action

Breaking out the egg-and-chicken problem of collecting ground-truth HRI data without actually having robots with mature social skills is either solved by remotely controlling

10.21437/AI-MHRI.2018-10

the robots by human pilots (so called *Wizard-of-Oz* or *teleoperation*) or by sketchy autonomous behavioral models. So the Vernissage corpus [Jayagopi *et al.*, 2013] comprises multiple auditory, visual, and robotic system information channels from the Nao robot while interacting with two persons as an art guide in a German art museum. The robot behavior (verbal output as well as gaze and nodding) was remotely controlled by a Wizard-of-Oz.

We expect here that the pilot provides the robot with optimal behaviors given the sensory information provided to him/her via the robot's sensors. But scaling human perception – or scene analysis performed by static sensors observing HHI – to active robots is not straightforward neither.

Several experiments have shown that multimodal signal processing is impoverished when performed by a moving platform, because of ego-noise, constant change of perspective ... Novoa *et al.* [2017] have shown that Word Error Rate (WER) – performed by a PR2 robot that moves its body and head while listening to sentences uttered by a fixed source – raises from 5.4% to 39.5% when displacement velocity is set to $0.6m/s$ and angular head rotation to $0.56rad/s$. For vision, few RGB-D datasets available today simulate robot motion through an environment [Ammirato *et al.*, 2017]. Separating impact of body motion (in particular when supporting sensors) from motions of objects and agents in the scene still remains a challenging problem. The majority of recent research employs motion information (via motor, proprioceptive or exteroceptive) to improve tracking and identification results [Rezazadegan *et al.*, 2017].

Note that current work makes use of passively collected ego-motion data [Agrawal *et al.*, 2015]. It remains to be seen if better multimodal representations are learned if the agent can actively decide on how to explore its environment (i.e. active learning [Bajcsy *et al.*, 2018] or interactive perception [Bohg *et al.*, 2017]).

3 Immersive teleoperation of robots

Development learning [Lee *et al.*, 2007] and learning by demonstration [Argall *et al.*, 2009] get around this retargeting problem by directly providing the robot with sensorimotor experiences. If these learning frameworks have been intensively explored for tasks involving contacts with the environment – such as walking, grasping, cooking ... – the field of HRI is more recent.

3.1 From Wizard-of-Oz experiments to immersive teleoperation

Most Wizard-of-Oz experiments in HRI [Riek, 2012] consist in asking one or several accomplices observing the HRI scene as third parties – often via semi-transparent mirrors – to trigger predefined verbal (assisting speech recognition or sentence generation) or non-verbal behaviors (such as performing head nods, pointing or gazing). The task of accomplices is mainly to guide decision and to decide *When* to act. When wizards do actually monitor actuators and sensors directly via their own body motions, they are preferably called *pilots* [Goodrich *et al.*, 2013]. When they perceive the robot's environment via robot's senses, this teleoperation is *immersive*.

One advantage of immersive teleoperation (also termed *beaming* [Normand *et al.*, 2012]) is to provide robots – whose social, emotional, linguistic as well as sensory motor capabilities are impoverished compared to human ones – with a cognitive control that supposedly takes the best use of available robotic features: a human brain *embodied* in a robot. This addresses two main issues already sketched in the introduction, that motivate the increasing interest in the training of social HRI by human demonstrations:

Scaling The pilot performs a intelligent sensorimotor mapping that is rightly scaled to the robot's morphology and dynamical abilities.

Human factors Human behaviors are monitored in a simulated HRI. Multimodal sensorimotor data collected by the robot during these passive experiences are very close to those that will be experienced during autonomous behaviors, if AI is able to reproduce the high-level cognitive behaviors that are implicitly recruited by the pilot.

These issues are mitigated both by technical and human factors: (a) the teleoperation platform should provide the pilot with high-quality sense of self-location, ownership and agency, with minimal cognitive overload – see in particular our effort for enabling faithful gaze control [Cambuzat *et al.*, 2018; Bailly *et al.*, 2018] and our teleoperation platform in fig. 1; (b) the pilot experience should also be augmented by the "superhuman" capabilities that are conversely expected from robots today – in particular in terms of episodic, autobiographic and encyclopedic memory – e.g. in contrast with conversational agents, pilots would not be able to instantaneously question the web to get recipes or the latest soccer scores. In that respect, our robot holds a tablet that enables the pilot to get instantaneous information about what the robot would have gathered from its information system and from the IoT. Symmetrically, note that Pepper also holds a screen on its torso in order to display multimedia information that are difficult to demonstrate via verbal and non verbal signals. From the pilot's (and somehow users') perspectives, the robot may be thus considered as a *cyber-physical gate* between virtual reality, IoT and HRI.

3.2 HRI and virtual reality

The area of Virtual, Augmented and Mixed Reality (VAMR) interactions between humans and robots – considering not only robots as a way to augment reality but also ways to perceive and act on cyber-physical spaces through an extended body – opens avenues for research and technology in the field of AI and HRI, by enabling humans and robots to share worlds, bodies and cognitions as well as"gamifying" manufacturing positions. Symptomatic of this trend are the recent workshops organized as satellite of key events of both communities:

Robotics and VAM the first International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interactions (VAM-HRI) was organized before HRI 2018 and the "Human in-the-loop robotic manipulation: on the influence of the human role" workshop at IROS 2017 that explored kinesthetic teaching and teleoperation.

VAM and Robotics The workshops "BCNAE: Body Con-

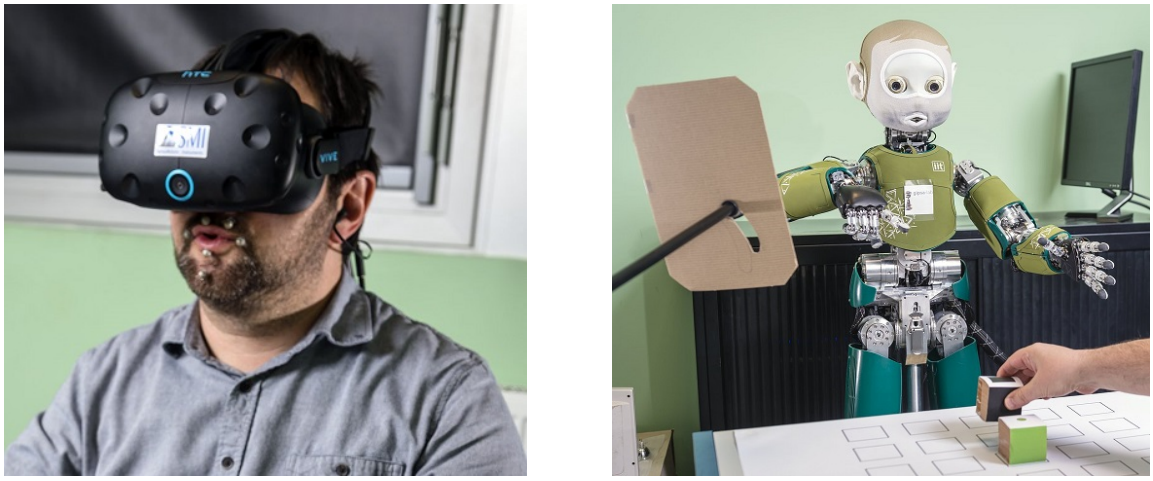

Figure 1: Beaming the head, eyes, lips and hands of Nina, the GIPSA-Lab iCub robot. Left, we re-target the head, eyes, lips and hand movements of the pilot to directly control the corresponding segments of Nina; The pilot receives audiovisual feedback from Nina's eye-embedded cameras and ear-microphones in the head-mounted display. Right: the remote HRI scene, where the teleoperated Nina instructs subjects to move objects. Note that instructions are given to the pilot via an augmented display that is overlaid onto the piece of cardboard. © Cyril FRESILLON / GIPSA-lab / CNRS Photothèque

sciousness in Natural and Artificial Environments" and "HAPTICS: Wearable and portable haptics for VR and AR" at IEEE VR 2018 explore the use of robotic systems in VAMR applications. ICRA 2018 will also host a workshop on "Robotics in Virtual Reality"

4 Modeling interactive multimodal behaviors

Generation of interactive multimodal behaviors of conversational agents often enriches a spoken dialog system that first manages verbal content and augments it with multimodal tags. One typical example is the SAIBA framework [Kopp *et al.*, 2006]: the Function Markup Language (FML) describes the agent's communicative functions that are further transformed into utterances tagged with micro-coordinated non-verbal behaviors described using the Behavioral Markup Language (BML). The action-perception loop was then closed by introducing a Perception Markup Language (PML) that converts input multi-sensory streams into stamped co-verbal events [Scherer *et al.*, 2012].

The advent of deep learning models that are capable of mapping multi-sensory input to semantic content (audiovisual speech, multimodal gesture vs. activity recognition, paralinguistic challenges aiming at estimating affects, physical or mental states of subjects) has changed the horizon. More recently, generative models have also been developed that conversely generate images, sounds and text from semantic content.

But AI *end-to-end* models able to directly map multi-sensory input streams – that should be aware of output actions and attentive to expected reactions – to motor streams are still scarce. If several works have demonstrated the efficiency of graphical models and deep learning in capturing causal relations between some multimodal signals in rather specific tasks (cf. backchannel opportunities [Ruede *et al.*, 2017], head movements [Ding *et al.*, 2014], gaze [Nguyen *et al.*, 2017b] ...) multimodal machine learning [Baltrušaitis *et al.*, 2018] still faces the problem of learning joint and coordinated representations that can be permeable to the task, the environmental conditions and adapt to the desired or observed *style* of interaction (see recent attempts for speech generation [Henter *et al.*, 2017; Wang *et al.*, 2018]).

5 Discussion

The use of AI for learning multimodal behavioral models for HRI still raises unsolved issues.

Interactive data vs. models Training an AI model that map multimodal perception to action with ground truth interactive data is a delicate challenge: since the model draw part of its input cues from reactions to actions that have not yet been performed, this odd open-loop training process may lead to unsuccessful expectations or unexpected reactions when used in a veridical close-loop test situation.

Domain of competence and social relevance If robots are all equipped with emergency red buttons, social misfunctioning is much harder to detect and process. First of all, it's hard to record adversarial social behaviors: by definition, humans deliver socially acceptable behaviors and non observed behaviors can either be false positive (non observed but acceptable variants) or just negative samples. One possibility for collecting negative samples is to ask perceivers to rate by consensus the model's behavior [De Kok *et al.*, 2010] when exploring unseen situations. Incidentally erroneous social behaviors could be then detected and penalized.

Evaluation Well-designed rewards and loss functions do not preclude subjective assessment that should go beyond self-assessments, behavioral measurements, psycho-physiological measures or task performance metrics [Sim and Loo, 2015]. We should be able to pro-

vide HRI engineers and designers with diagnostic tools that can identify *What* and *When* social behaviors or co-adaptation went wrong. Localized HRI events – lack of responsiveness, improper social signals ... – can in fact strongly degrade subjective evaluation of behaviors despite a better goodness-of-fit. New evaluation methodologies should be proposed, in particular that give access to on-line processing of HRI by involved subjects or third parties (e.g. see [Nguyen *et al.*, 2017a]).

References

[Agrawal *et al.*, 2015] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Int. Conf. on Computer Vision (ICCV)*, pages 37–45. IEEE, 2015.

[Ammirato *et al.*, 2017] Phil Ammirato, Patrick Poirson, Eunbyung Park, Jana Košecká, and Alexander C Berg. A dataset for developing and benchmarking active vision. In *Int. Conf. on Robotics and Automation (ICRA)*, pages 1378–1385. IEEE, 2017.

[Argall *et al.*, 2009] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[Azagra *et al.*, 2017] Pablo Azagra, Florian Golemo, Yoan Mollard, Manuel Lopes, Javier Civera, and Ana Murillo. A multimodal dataset for object model learning from natural human-robot interaction. In *Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 6134–6141. IEEE/RSJ, 2017.

[Bailly *et al.*, 2018] Gérard Bailly, Chrsitian Wolf Alaeddine Mihoub, and Frédéric Elisei. Gaze and face-to-face interaction: from multimodal data to behavioral models. In Gert Brône and Bernt Oben, editors, *Advances in Interaction Studies. Eye-tracking in interaction. Studies on the role of eye gaze in dialogue*. John Benjamins, Amsterdam, 2018.

[Bajcsy *et al.*, 2018] Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. Revisiting active perception. *Autonomous Robots*, 42(2):177–196, 2018.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[Bilakhia *et al.*, 2015] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61, 2015.

[Bohg *et al.*, 2017] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017.

[Cambuzat *et al.*, 2018] Rémi Cambuzat, Frédéric Elisei, Gérard Bailly, Olivier Simonin, and Anne Spalanzani. Immersive teleoperation of the eye gaze of social robots. In *Int. Symposium on Robotics (ISR)*, 2018.

[Castellano *et al.*, 2010] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W McOwan. Inter-act: An affective and contextually rich multimodal video corpus for studying interaction with robots. In *Int. conf. on Multimedia*, pages 1031–1034. ACM, 2010.

[De Kok *et al.*, 2010] Iwan De Kok, Derya Ozkan, Dirk Heylen, and Louis-Philippe Morency. Learning and evaluating response prediction models using parallel listener consensus. In *Int. Conf. on Multimodal Interfaces and Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI)*, pages 3:1–3:8. ACM, 2010.

[Ding *et al.*, 2014] Chuang Ding, Pengcheng Zhu, Lei Xie, Dongmei Jiang, and Zhong-Hua Fu. Speech-driven head motion synthesis using neural networks. In *Interspeech*, pages 2303–2307. ISCA, 2014.

[Goodrich *et al.*, 2013] Michael A Goodrich, Jacob W Crandall, and Emilia Barakova. Teleoperation and beyond for assistive humanoid robots. *Reviews of Human factors and ergonomics*, 9(1):175–226, 2013.

[Henter *et al.*, 2017] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. Principles for learning controllable tts from annotated and latent variation. In *Interspeech*, pages 3956–3960. ISCA, 2017.

[Jayagopi *et al.*, 2013] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede, Vasil Khalidov, Laurent Nyugen, Britta Wrede, and Daniel Gatica-Perez. The vernissage corpus: A conversational human-robot-interaction dataset. In *Int. conf. on Human-robot interaction (HRI)*, pages 149–150. IEEE, 2013.

[Kendon, 2004] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[Kita, 2003] Sotaro Kita. *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003.

[Kopp *et al.*, 2006] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R Thórisson, and Hannes Vilhjálmsson. Towards a common framework for multimodal generation: The behavior markup language. In *Int. workshop on intelligent virtual agents (IVA)*, pages 205–217. Springer, 2006.

[Krenn and Pirker, 2004] Brigitte Krenn and Hannes Pirker. Defining the gesticon: Language and gesture coordination for interacting embodied agents. In *AISB Symposium on Language, Speech and Gesture for Expressive Characters*, pages 107–115, 2004.

[Lee *et al.*, 2007] Mark H Lee, Qinggang Meng, and Fei Chao. Developmental learning for autonomous robots. *Robotics and Autonomous Systems*, 55(9):750–759, 2007.

[McKeown *et al.*, 2012] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2012.

[Nguyen *et al.*, 2017a] Duc-Canh Nguyen, Gérard Bailly, and Frédéric Elisei. An evaluation framework to assess and correct the multimodal behavior of a humanoid robot in human-robot interaction. In *GEstures and SPeech in INteraction (GESPIN)*, pages 56–62. ISCA, 2017.

[Nguyen *et al.*, 2017b] Duc-Canh Nguyen, Gérard Bailly, and Frédéric Elisei. Learning off-line vs. on-line models of interactive multimodal behaviors with recurrent neural networks. *Pattern Recognition Letters*, 100:29–36, 2017.

[Normand *et al.*, 2012] Jean-Marie Normand, Maria V Sanchez-Vives, Christian Waechter, Elias Giannopoulos, Bernhard Grosswindhager, Bernhard Spanlang, Christoph Guger, Gudrun Klinker, Mandayam A Srinivasan, and Mel Slater. Beaming into the rat world: enabling real-time interaction between rat and human each at their own scale. *PloS one*, 7(10):e48331, 2012.

[Novoa *et al.*, 2017] José Novoa, Jorge Wuth, Juan Pablo Escudero, Josué Fredes, Rodrigo Mahu, Richard Stern, and Nestor Becerra Yoma. Robustness over time-varying channels in dnn-hmm asr based human-robot interaction. In *Interspeech*, pages 839–843. ISCA, 2017.

[Oertel *et al.*, 2014] Catharine Oertel, Kenneth A Funes Mora, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. Who will get the grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32. ACM, 2014.

[Rezazadegan *et al.*, 2017] Fahimeh Rezazadegan, Sareh Shirazi, Ben Upcrofit, and Michael Milford. Action recognition: From static datasets to moving robots. In *Int. Conf. on Robotics and Automation (ICRA)*, pages 3185–3191. IEEE, 2017.

[Riek, 2012] Laurel D Riek. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136, 2012.

[Ruede *et al.*, 2017] Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. Enhancing backchannel prediction using word embeddings. In *Interspeech*, pages 879–883. ISCA, 2017.

[Scherer *et al.*, 2012] Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini, Alesia Egan, Louis-Philippe Morency, et al. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *Int. Conf. on Intelligent Virtual Agents (IVA)*, pages 455–463. Springer, 2012.

[Sim and Loo, 2015] Doreen Ying Ying Sim and Chu Kiong Loo. Extensive assessment and evaluation methodologies on assistive social robots for modelling human–robot interaction–a review. *Information Sciences*, 301:305–344, 2015.

[Wang *et al.*, 2018] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.