Embodied Multi-modal Interaction in Language learning: the EMIL data collection

Stefan Heinrich, Matthias Kerzel, Erik Strahl, Stefan Wermter

Knowledge Technology, Department of Informatics, Universität Hamburg, Germany
{heinrich,kerzel,strahl,wermter}@informatik.uni-hamburg.de
http://www.knowledge-technology.info

I. Introduction

Humans develop cognitive functions from a body-rational perspective. Particularly, infants develop representations through sensorimotor environmental interactions and goal-directed actions [1]. This embodiment plays a major role in modeling cognitive functions from active perception to natural language learning. For the developmental robotics community, working with humanoid robotic proxies, datasets are interesting that provide low-level multi-modal perception during the environmental interactions [2].

Related Data Sets: In the last years, many labs made considerable efforts to provide such datasets, focussing on different research goals but also taking technical limitations into account. Examples include: the KIT Motion-Language set for descriptions of whole-body poses [3], the MOD165 set of a gripper-robot having vision, audio, and tactile senses for interacting with objects [4], the Core50 set focussing on human perspective and vision [5], and the similar but up-scaled EMMI and iCubWorld sets [6]. However, none of these corpora provide true continuous multi-modal perception for interaction cases, as we would expect an infant is experiencing.

In this preview, we introduce the Embodied Multi-modal Interaction in Language learning (EMIL) data collection, an ongoing series of datasets for studying human cognitive functions on developmental robots. Since we aim to utilize resources in tight collaboration with the research community, we propose the first set on object manipulation for fostering discussions on future directions and needs within the community[1].

II. Dataset Characteristics

In this first set, the developmental robot NICO is mimicking an infant that interacts with objects and receives a linguistic label after an interaction. The interaction follows usual interaction schemes of 12–24 month-old infants on toy-like objects.

Developmental Robot Setup: In developmental robotics, the goal is to study human cognitive functions in conditions of human infants interacting in natural environments [2]. These conditions include *embodied* interaction with natural motor and sensing capabilities of an infant and multi-modal sensations within active perception [7]. For our data recording, we developed a child-like humanoid robot and utilize it in scenarios that resemble natural infant environments, such as in playing with objects at a table.

Interactive Robot NICO: Our developmental robot is the Neuro-Inspired COmpanion *NICO* [8], which is mainly aimed for research on multi-modal human-robot interaction and neuro-cognitive modelling. NICO includes two HD RGB cameras, stereo auditory perception, and interaction capabilities of a 3.5-year-old child, including six DOF arms and hands with multi-grip fingers and tactile sensors. In interactions with objects, the robot's hands as well as the whole upper body provide perception on a sensorimotor level and at the same time introduce the interaction imprecision and self-occlusion in a way our infants show.

Recording: In the setup, NICO is seated in a child-sized chair at a table, interacting with the right hand and facing the head downwards during the experiment, while a human places a small object on the table at a fixed position (see Fig. 1). A predefined action is carried out on the object, e.g., lifting it up or scooting it across the table. During the robot's actions, a continuous multi-modal recording encompasses continuous streams of visual information from the left and right robot camera as well as from the external experimenter, stereo audio information, and proprioceptive information from the robot's body. Finally, the experimenter provides a linguistic label.

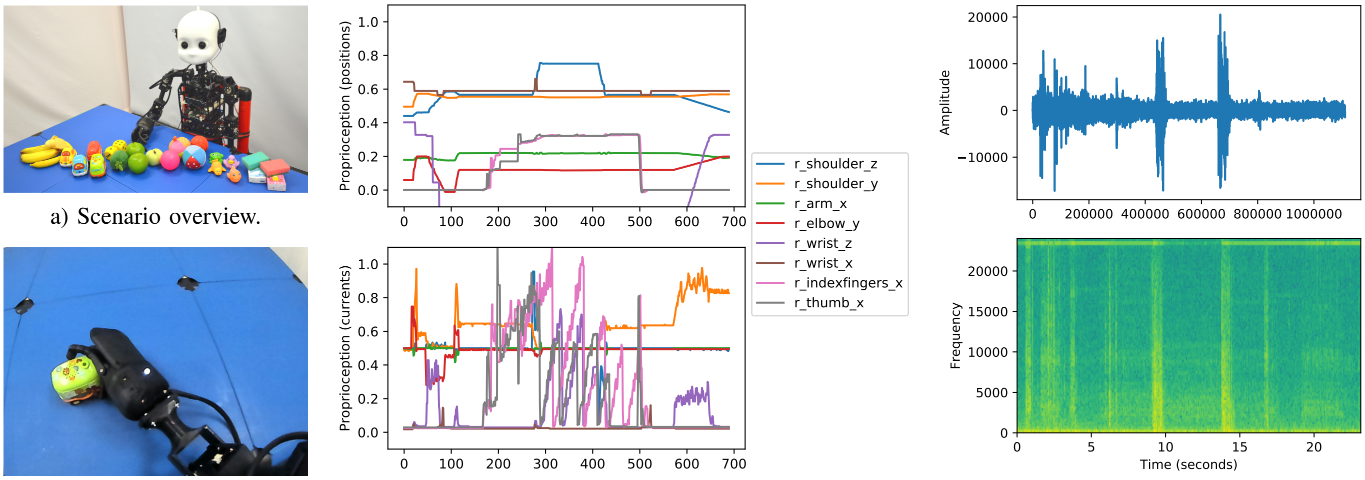
III. Impact and Research Opportunities

Our continuous, multi-modal, and particularly body-rational data allows for studying a large range of algorithms on fundamental classification or prediction tasks. This includes object recognition and tracking, action recognition, and question answering. Moreover, the dataset is aimed at research in a range of state-of-the-art research topics.

Active Perception: The different actions and objects allow to build up a training scheme within a model by selecting to experience a certain interaction because the model estimates that this provides the highest information gain or reduces uncertainty. In humans, we find the tendencies that a perception choice or a specific action is voluntary [9]. Thus, the dataset is suited for developing models that aim to explain how the sensory input gathered from an object with different, multi-modal sensors changes based on the robot's actions.

The authors gratefully acknowledge partial support from the German Research Foundation DFG under project CML (TRR 169).

[1] The EMIL data collection will be made on several stages available via: http://corpora.knowledge-technology.info



a) Scenario overview.

b) Visual perception (right).

b) Sensorimotor perception.

c) Auditory perception.

Fig. 1: Characteristics of the EMIL dataset: Continuous multi-modal recording of interactions with objects.

Cross-modal Representation Learning: Since the different recorded modalities include information about the same object and interaction quite differently, the dataset is suited to study algorithms on multi-modal and cross-channel representation learning. For some objects and actions the data contains salient features in a certain modality, while for others, all modalities are necessary for disambiguation. This allows studying mechanisms on sensor fusion, superadditivity, and hierarchical composition in addition to embodied representation formation on the cortex-level [10].

Developmental Language Acquisition: A research question related to representation learning is natural language acquisition since representations for language production and language perception in the human brain seem to form embodied and cross-modally integrated [1], [2]. The dataset is therefore particularly suited for research on the grounding of language in sensorimotor perception because the recording diligently followed the developmental robot approach [11]. Mechanisms for representation formation and bidirectional hierarchical composition and decomposition can get tested in the biologically plausible setting.

As a second step, this allows extending this dataset by much larger parts of abstract and ungrounded linguistic input, in a fashion that parents would provide verbally or with the aid of a storybook to their infant [12]. Here, language acquisition models can get studied for how they integrate additional knowledge into their grounded representations, but also how a teaching application can provide suitable teaching content.

Livelong Learning: The dataset is suited to provide evaluation data for (neural) lifelong learning approaches [13]. An initial subset of the training data can be selected that is limited to a few types of objects, actions or just a low number of samples. Over the course of time, life-long learning experiences can be simulated by adding more and more parts of the data-set to the learning.

IV. Conclusion

The proposed data collection EMIL aims at providing researchers from the developmental robotics and related fields the opportunity to research into intriguing questions around human cognitive functions. For the workshop, we invite researchers to jointly develop the future data sets to come.

References

[1] S. Heinrich and S. Wermter, "Interactive natural language acquisition in a multi-modal recurrent neural architecture," *Connection Science*, vol. 30, no. 1, 2018.

[2] A. Cangelosi and M. Schlesinger, *Developmental robotics: From babies to robots*. Cambridge, US: The MIT Press, 2015.

[3] M. Plappert, C. Mandery, and T. Asfour, "The KIT motion-language dataset," *Big Data*, vol. 4, no. 4, pp. 236–252, 2016.

[4] T. Nakamura and T. Nagai, "Ensemble-of-concept models for unsupervised formation of multiple categories," *IEEE Transactions on Cognitive and Developmental Systems*, vol. online, 2017.

[5] V. Lomonaco and D. Maltoni, "Core50: a new dataset and benchmark for continuous object recognition," in *Proc. PMLR*, 2017, pp. 17–26.

[6] X. Wang, F. M. Eliott, J. Ainooson, J. H. Palmer, and M. Kunda, "An object is worth six thousand pictures: The egocentric, manual, multi-image (EMMI) dataset," in *Proc. ICCV WS*, 2017, pp. 2364–2372.

[7] J. Tani, *Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena*. Oxford University Press, 2016.

[8] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "NICO - Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction," in *Proc. IEEE RO-MAN*, 2017, pp. 113–120.

[9] P.-Y. Oudeyer, "Computational theories of curiosity-driven learning," *arXiv preprint*, 2018, https://arxiv.org/abs/1802.10546.

[10] J. Bauer, J. Dávila-Chacón, and S. Wermter, "Modeling development of natural multi-sensory integration using neural self-organisation and probabilistic population codes," *Connection Science*, vol. 27, no. 4, 2015.

[11] C. Lyon, C. L. Nehaniv, J. Saunders, T. Belpaeme, A. Bisio, K. Fischer, F. Förster, H. Lehmann, G. Metta, V. Mohan *et al.*, "Embodied language learning and cognitive bootstrapping: methods and design principles," *International Journal of Advanced Robotic Systems*, vol. 13, no. 3, 2016.

[12] S. Heinrich, C. Weber, S. Wermter, R. Xie, Y. Lin, and Z. Liu, "Crossmodal language grounding, learning, and teaching," in *Proc. CoCo@NIPS2016)*, 2016, pp. 62–68.

[13] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *arXiv preprint*, 2018, https://arxiv.org/abs/1802.07569.