2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

From Human to Robot Everyday Activity

Celeste Mason[1], Konrad Gadzicki[2], Moritz Meier[1], Florian Ahrens[3], Thorsten Kluss[2], Jaime Maldonado[2], Felix Putze[1], Thorsten Fehr[3], Christoph Zetzsche[2], Manfred Herrmann[3], Kerstin Schill[2], Tanja Schultz[1]

***Abstract*— The Everyday Activities Science and Engineering (EASE) Collaborative Research Consortium's mission to enhance the performance of cognition-enabled robots establishes its foundation in the EASE Human Activities Data Analysis Pipeline. Through collection of diverse human activity information resources, enrichment with contextually relevant annotations, and subsequent multimodal analysis of the combined data sources, the pipeline described will provide a rich resource for robot planning researchers, through incorporation in the OpenEASE cloud platform.**

I. INTRODUCTION

Currently, robots have displayed remarkable feats that would suggest they will soon be able to take over many of our more onerous daily activities (cleaning, cooking, feeding the dog etc), leaving us free to focus our energies elsewhere (eating, petting the dog etc). However, the underlying truth remains – these robots display such sophisticated abilities due to their creators' contextually precise, expertly crafted planning algorithms. For this everyday robotic revolution to occur, these agents will need to be able to react to vague instructions and changing context, in a manner that more closely adheres to human behaviors and abilities. So, how may we identify and, more importantly, collect and describe the missing pieces of this puzzle that would enable cognitive robots to perform actions that approach the aplomb with which humans are able to interact everyday, through habit, common sense, intuition, and problem solving approaches seemingly effortlessly developed throughout their lifetimes?

In this paper we present a novel data processing pipeline for human activity recognition (HAR). To our knowledge, our pipeline is the first to combine multimodal data collection, hierarchical and semantic annotations, and ontological reasoning to enhance cognitive robots with human-like reasoning capabilities derived from everyday human activities. The collaborative research center EASE ("Everyday Science and Engineering," http://ease-crc.org) has facilitation of robotic mastery of everyday activities as its unifying mission. The subprojects concerned with human activities data collection have the goal of providing so-called narrative-enabled episodic memories (NEEMs). These data structures store recorded observations, experiences and activities compiled as a single coherent item. Another goal is the derivation of pragmatic everyday activity manifolds (PEAMs), which will form the basis for robot agent enhancement by enabling real-time interaction similar to how humans function.

[1]Cognitive Systems Lab, University of Bremen, Germany
[2]Cognitive Neuroinformatics, University of Bremen, Germany
[3]Neuropsychology and Behavioral Neurobiology, University of Bremen, Germany

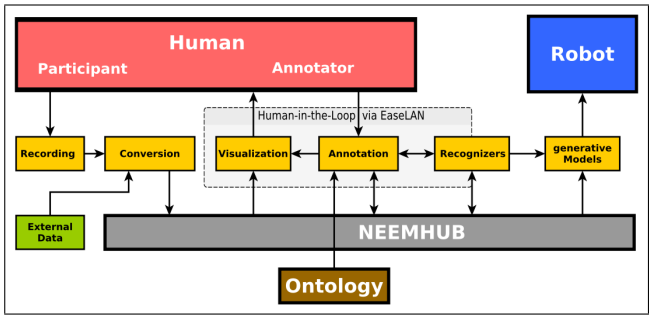

Fig. 1: The EASE human activities data analysis pipeline

We start with recording of human activities in a kitchen setting, preprocessing, and (optionally) supplementing with outside data sources, before storing data in the openEASE robotics knowledge base platform. Annotations, based on the EASE-Ontology, are designed for cognitive robots. The automatic annotators use different modalities and approaches, e.g. multimodal activity recognition, speech recognition or object tracking, thus complementing each other. Data produced from further processing through manual and automatic annotation, and subsequent analyses through a variety of machine learning techniques, can then be queried in openEASE. Based on performance in robotic activity scenarios, the annotation schema can then be improved further. The results–raw and processed multimodal data recordings, together with annotations and data derived from analyses–are stored in openEASE, a framework for knowledge representation and reasoning, as shown in figure 1.

The NEEMs derived from research projects in EASE subproject area H (Human Activities Data Collection) provide unique and critical contextual background for robots, based on human activities, perceptions, and feedback. Analyses of biosignals derived from brain, muscle, or skin signals may provide insight into diverse aspects of human behavior required for humans to masterfully perform everyday activities with little effort or attention. Through the integration of analyses from a wide array of data sources, through a multitude of complementary methods, we endeavor to build an extensive, contextually dense reserve of activity experience and problem solving approach methods derived from human behaviors to transfer effectively to robot systems. The following examples are among those being employed within the EASE-CRC at this time.

Brain activity measurements allow evaluation of attentional focus while performing tasks, adaptation to ambiguous or conflicting situations or physical obstacles, decision

making processes, and how motor imagery when viewing performance of activities compares with in-situ motor execution. Skin and muscle activity sensors can indicate overall mental state, and information about manual manipulation interactions with objects, such as the force used. Full body motion capture provides motion in an environment and object interactions. Assessment of small scale hand movements (including e.g. forces, velocities, trajectories, etc.) using the PHANToM haptic interface Scene video from many perspectives allows tracking of objects and the order of interactions, insight into efficient movement within a space while performing tasks. First person video provides understanding of scene aspects people may focus on while planning and executing tasks. Important information for a robot might include attention (internal vs external) and visual search strategies for objects or positions based on contexts such as meal type, formality, or number of diners. Microphones record scene audio, speech and non-speech vocalizations. Through audio recordings of what a person thinks-aloud while they perform tasks, we gain a rich description of the scene as the performer sees it, obstacles encountered, reasoning and problem solving approaches, frustration or enjoyment, and the task process as a whole.

openEASE [1] is an online knowledge representation and reasoning framework for robots. It provides the infrastructure to store and access nonhomogeneous experience data from robots and humans, and comes with software tools which enable researchers to analyze and visualize the data.

EASE subprojects record human activity datasets (HAD) in a range of scenarios. Data collection efforts focus on contexts involving "Activities of Daily Living" (ADLs) in the kitchen, such as setting and clearing the table or doing dishes. From these experiments, we are producing the multimodal EASE Table Setting Dataset (EASE-TSD), a dataset featuring brain measurements using functional magnetic resonance imaging (fMRI) and electroencephalography (EEG) during table setting related tasks, and the EASE Manipulation Adaptivity Dataset (EASE-MAD), that focuses on sensorimotor regulation during individual (atomic) actions or short sequences, in detail. These experimental contexts and data recordings are described in later sections.

Seamless integration of these datasets, annotations, and derived models into the openEASE framework will provide the solid foundation for robotic researchers to expedite development of robotic agents that are more robust to unexpected variations in task requirements and context, taking human behaviour as inspiration.

II. RELATED WORK

A. Data for Activity Recognition

Activity recognition can be performed on a wide variety of features and a large number of datasets have been provided (for a review see [2]).

Recent approaches in activity recognition often work on RGB-D data. These consist of RGB video and accompanying depth maps and provide two useful modalities for human activity recognition. Skeleton data can be recorded with motion capturing or extracted from RGB-D data as in the case of a Kinect. Other modalities, e.g. optical flow, can be extracted from RGB data as well.

There are several other datasets featuring kitchen related activities. "EPIC-Kitchens" [3] features head-mounted camera video taken during kitchen activities performed by 32 participants in their homes, annotated with 125 verb classes and 352 noun classes in varied languages. "50 Salads" [4] uses 3rd-person camera (including RGB-D cameras) and accelerometer-mounted objects to record meal preparation sequences. "MPII Cooking Activities Dataset" [5] uses video recordings of participants performing 65 kitchen activities for pose-based activity recognition. The "TUM Kitchen Dataset" [6] features video, full-body motion capture, RFID tag readings and magnetic sensor data taken during activities, processed with manual motion tracker labels and automatic semantic segmentation.

B. Activity Recognition Models with Neural Networks

The specific properties of the various modalities have led to different processing strategies. For instance, the RGB channel can be processed with a spatio-temporal convolutional neural network (CNN) [7], [8], [9], [10], either on its own or together with derived optical flow [11], [12] through a two-stream CNN or as a multistage CNN [13], [14]. Another approach is to use recurrent neural networks (RNNs) for processing of RGB data [15], [16], [17], [18], [19]. The depth channel can be similarly processed with a CNN [20], [21] or with a combination of CNN and RNN [22]. With regard to skeleton data, processing with CNN can be enabled by interpreting joint positions as image data [23], [24], [25], [26]. There also exist RNN-based approaches [27], [28], [29], a Deep Boltzmann Machine (DBM) approach [30] and a Hidden Markov Model (HMM) with a deep network as a state probability predictor [31].

C. Semantic, Multimodal Activity Recognition

In [32], semantic hierarchically structured actions are recognized within the kitchen-related context of pancake making, sandwich making, and setting the table in order to transfer task-related skills to humanoid robots. Their ontologically-associated knowledge representations of the observed behavioral data, recorded as video, of people during interactions with objects during such tasks is defined at varying levels of abstraction. Semantic activity recognition of kitchen ADLs (such as making pasta or taking medicine) in the form of multimodal sensor data [33] has also been performed, supported by a Semantic Sensor Network ontology for worn and environment sensor information.

III. DATA RECORDING

A. Human Activities in a Pseudo-natural Setting

The EASE Table Setting Dataset (EASE-TSD) is composed of multimodal biosignal data recorded during experimental observations of various table setting tasks performed by participants in our Biosignals Acquisition Space (EASE-BASE), as described in [34]. All signals are recorded

synchronously using Lab Streaming Layer (LSL) [35]. The recorded sensor modalities include: full-body motion-tracking, audio (from a scene mic and head-worn mic for speech), video from 7 mounted webcams and one head-mounted eyetracker, and biosignals from muscle and brain activity) from participants performing everyday activities while describing their task through use of think-aloud protocols during the task (concurrently) and after the task is completed (retrospectively) [36], as shown in Figure 2a-b. For the EASE-TSD experimental recordings, 70 sessions have been recorded, composed of six or more trials each, totaling 470 concurrent and 405 retrospective think-aloud trial variants. Over 37,400 transcribed words of the think-aloud speech during these trials have been created, with think-aloud encoding annotations underway. Over 16,600 action annotation labels, broken down into between 2 and 12 category sets, for these trials have been performed at varying levels of granularity. Annotations and transcriptions on numerous levels continue to be created as analyses progress. Once the planned recordings with 100 participants are finished and preliminary analysis is completed, the data set will be made available to the public.

Fig. 2: Experimental data recording of participants a) performing concurrent think-aloud trial tasks, b) performing retrospective think-aloud trial tasks c) performing tasks during fMRI, and d) performing tasks during stationary EEG.

B. Human Activities in a Controlled Setting

Table setting videos recorded from the first-person perspective are used in neuroimaging studies, using a 3-Tesla MRI-Scanner as well as high-density multi-channel EEG system, situated in an electromagnetically shielded room. Study participants are tasked with actively imagining themselves acting out the presented situations, thus employing motor imagery [37], while their brain activity is measured.

While fMRI offers unrivaled spatial resolution and the ability to accurately measure whole brain volumes, the residential EEG-System provides high temporal resolution and, in comparison to mobile solutions, offers the advantage of less data contamination caused body movement and electromagnetic emission sources such as cameras, movement detection systems and so forth. Due to its high number of acquisition channels, it also allows for detailed source localization of brain activity.

C. Adaptivity of Human Activities

The Manipulation Adaptivity Dataset (EASE-MAD) includes data from different sources that were assessed in controlled VR settings. This approach allows for deeper insights into the *sensorimotor loop (SML)*, which is a model concept for describing the integration of sensory and motor systems that is the basis of continuous modification of motor commands in response to sensory inputs. Whereas basic sensorimotor loops have been successfully modeled in several variants using control engineering approaches [38], they cannot explain the effortless precision and vast flexibility found in human voluntary actions.

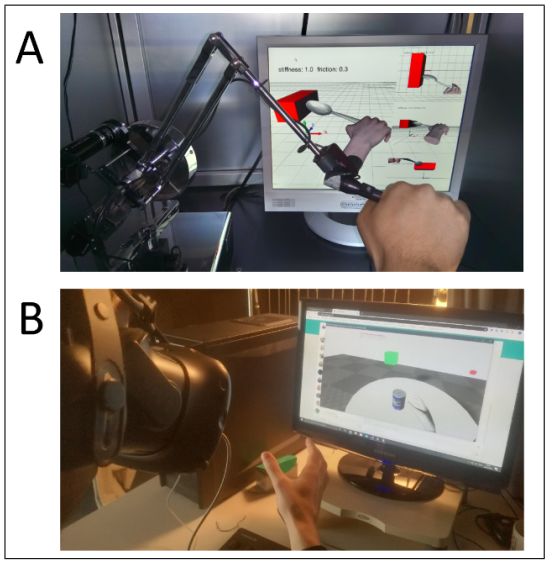

Fig. 3: Data recording during adaptivity testing a) using the PHANToM haptic interface that allows for force rendering to interact with objects and tools in VR, and b) using optical tracking to control a full hand model in VR. The head mounted display is equipped with an eyetracker.

It should be noted, however, that observation of real-world everyday activities only allows us to capture the sensorimotor loop from the outside (i.e. analyze its *outer* PEAMs without being able to directly address its inner laws. We seek more direct access by closing the sensorimotor loop in *virtual reality* (VR), in which we are, unlike in real world experiments, in full control of the individual parameters of the environment and actions. Thus, our main research paradigm will be to intervene in the sensorimotor loop at different points of the control chain [39], [40]). VR as experimental setting enables systematic intervention beyond the physical limitations of real world studies. This allows analysis of how cognitive systems adjust to changing and ambiguous environmental conditions and a systematic modeling of both the inner and outer PEAMs of everyday activities.

Figure 3 shows experimental setups to record multimodal data including e.g. grasping trajectories, hand pose and finger positions during an action, or applied forces (assessed with an PHANToM haptic interface (e.g. [41], [42]). For a detailed description of data acquisition methods for the EASE-MAD

Dataset and the underlying approach to investigate human sensorimotor adaptivity see [43], [44], [45].

IV. DATA ANNOTATION

A. Annotation Schema and Ontology Integration

Annotation and transcription schema developed for the EASE-TSD to describe aspects of everyday activities pertinent to robotic planning algorithm improvement, and are therefore aligned to the EASE Ontology, are used to annotate video and transcribe audio from speech recorded during the EASE-TSD trials. The annotation schema for video-based recordings are hierarchically-structured semantic descriptions of events at increasingly fine-grained levels of detail. The highest level is the task phase (planning, object retrieval, etc.). Below that are specific recurring action types (picking up objects, searching for places to set them on the table). Actions are further broken down into motions (picking = reach, grasp, lift, retract arm) for each hand or other differentiating criteria. When multiple actions occur simultaneously, multiple annotation tiers are required.

B. Annotation Process

Annotation of various modalities is performed in accordance with the requirements for each type of data, first manually then through automated processing. For the EASE-TSD, the annotation and transcription processes are primarily performed in ELAN, as seen in Figure 4. For each trial, annotation or transcription is performed by one person, then checked by a second. As more data is collected and annotated, additional annotations will continually be performed by additional annotators on previously annotated data, then followed by inter-rater reliability scoring.

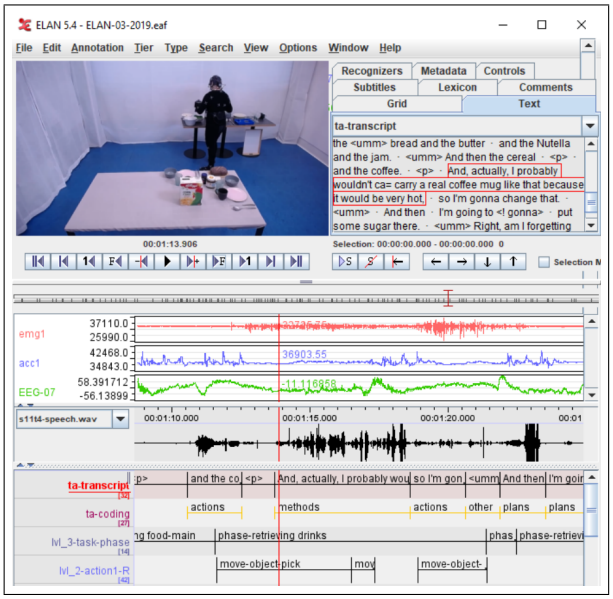

Fig. 4: ELAN is used to create transcriptions and annotations from audiovisual and biosignal data.

Video from multiple angles is used to label time segments where a person is performing specific actions. Frames from these videos are used to obtain information about the number, layout, and positioning of objects within the scene.

Transcription of speech is performed for both concurrent and retrospective think-aloud protocols, with the final goal of transcription by at least two transcribers. These think aloud trials are then coded based on an utterance level schema, to describe the types of thought processes and topics each participant thought relevant to the task at hand at the time.

V. ANALYSIS

A. Human Activity Recognition with Multimodal CNNs

Within our pipeline, we would like to automatically annotate our data. For annotation of video, we used a multimodal fusion approach based on CNNs.

Multimodal fusion is a popular approach to increase the performance of a machine learning system by using several data jointly. It comes in different flavors, with early, late and hybrid fusion being the primary distinctive types [46], with the main difference being where in the processing chain the fusion takes place. All those different types come with different advantages and challenges. The most simple fusion is probably late fusion [47]. Here each modality is processed separately and results are fused afterwards. It allows for maximum flexibility in choosing the processing method for each modality, so that one could use sophisticated unimodal systems (e.g. classifiers) and combine their outputs by i.e. summation, averaging or majority vote. It lacks the potential to exploit possible cross-correlations which may exist between the different data. Early fusion [47] offers a way to exploit those. Here, either raw data or data produced by feature extraction are fused in the beginning of the processing, in the most simple case by concatenation. Afterward the combined data are processed together. This approach requires that the input data are aligned, which might not be trivial when one has to deal with different dimensionality, sampling rate etc. Furthermore there is no choice of specialized approaches for separate modalities; the chosen approach has to fit the joint data.

We have developed a system which allows for the fusion at arbitrary layers. We define a splitting point within the network, up to which the different modalities are processed separately. Afterwards the merged layers are processed by the remaining network. The underlying architecture of our CNN is a “Kinetics I3D” [12] which uses “Inception 3D” units for spatio-temporal processing.

We have used an early fusion CNN approach for this work since in [48] we could show that for activity recognition early fusion performs better than late fusion. For early fusion the individual modalities are processed by the first convolutional layer separately. Its results are fused by concatenation for further processing. Figure 5 shows our architecture for early fusion. Apart from the fusion step, it is a standard “Kinetics I3D” implementation in TensorFlow [49] with sparse softmax cross entropy for loss calculation during training.

Based on the RGB videos, optical flow has been computed with “FlowNet2” [50]. The original full HD RGB videos were rescaled and cropped to 224x224 pixels, the same has

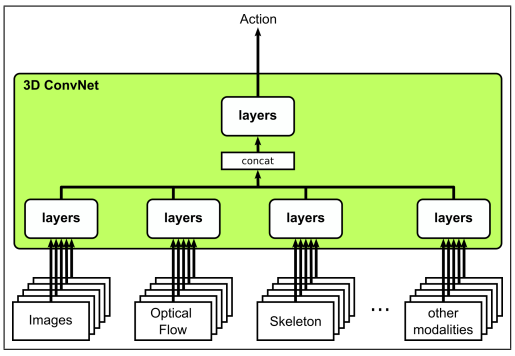

Fig. 5: Early fusion of multiple modalities in a CNN. The modalities are processed up to a specific point in individual paths, then fused by concatenation for further processing.

been done with the optical flow videos based on the RGB videos. See [48] for details of the multimodal network.

The EASE-TSD videos are typically several minutes long. To process them, the data has been chunked into slices during pre-processing. For each slice, we extract the associated ground truth labels which are present during the time slice. Since there can be more than one active label in the time span of a time slice, we have enabled our system to produce multi-label outputs as well as single labels. The multi-label training uses binary cross-entropy as the loss function while the single label variant uses categorical cross-entropy.

Depending on the granularity of the activities within the hierarchy, different time slices might be most useful. For low level actions like `reach` or `pick`, a brief time window of $250ms$ might be sufficient, while a higher level composed action like `pick & place` might not be properly recognized, requiring a longer slice of up to $2s$.

B. Speech Processing

Alongside the human-data table setting recordings, verbal reports of the performed actions and thought processes are recorded. As soon as a new speech recording is present, human transcribers are assigned to a transcription process, which is specified by explicit transcription rules. Automatic speech recognition with pretrained Kaldi acoustic models trained on the GlobalPhone corpus together with a language model enhanced by the previous transcripts is employed to aid the transcription process. A custom ELAN plugin generates a transcript with fine grained segments.

C. Multimodal Biosignal Action Recognition

The analysis of fMRI-data is based on different methodological approaches. Statistical models such as the General Linear Model (GLM) and Independent Component Analysis (ICA) allow contrastive analysis of differences in spatiotemporal patterns of brain activity related to annotated semantic episodes within a perceived point in time during video presentation (e.g. pick up, place, carry), thus leading to detection of distinct neuronal networks correlating with ontological categories. In particular, focus will be on the analysis of the level of neuronal network complexity during planning and execution of complex everyday activities. NEEMs and PEAMs generated from these categorized episodes of brain activity will then be contributed to openEASE.

Building on this knowledge of brain areas that are closely correlated in their activity to ontological categories, further research will also aim at developing algorithms that predict stimuli and semantic episodes on different levels of complexity. Thus, a semi-automatic scene recognition approach will be developed which can feed information into the planned process of automatic activity recognition and its annotation.

Furthermore, the combined use of multi-channel EEG and fMRI allows for a detailed examination of spatiotemporal characteristics of event-related brain activity [46] by using the spatial activation patterns derived from the analysis of the fMRI-data as seed regions for fMRI-constrained source analyses of EEG data [47]. EEG data will be first examined via Fourier analyses (FFT) and band-pass filtered according to oscillatory specificities of ontologically different time periods identified by topographical signal space analyses. Source analyses techniques will then be applied to determine characteristics of the spatial distribution and the spatiotemporal complexity of different periods of table setting action.

For EASE-TSD biosignal-based action recognition, we work toward a model to decode arm movements involved in object manipulation during typical table setting tasks from brain and muscle activity signals, captured by mobile electroencephalography (EEG) and electromyography (EMG) sensors. A subset of 50 EASE-TSD trials performed by 15 participants, manually annotated at the lowest level of arm motion, were used as the basis for multi-class classification using a convolutional and long-short term memory (CLSTM) model on spatially and temporally extracted features derived from EEG and EMG data. This subset of data was recorded using mobile EEG using 16 channels on the scalp as well as 4 EMG sensors each placed on the forearms. After undergoing channel selection, preprocessing, and then statistical and spatiotemporal feature extraction, this became the basis for classification of pick and place activities. For this experimental scenario from the EASE-TSD, we used recordings from experiments performed by 15 right handed participants, 8 male, with age ranging from 20 to 30 as described in [51].

At the lowest level, data from sensors placed at four positions on each arm (e.g., on muscles controlling hand activity of the right forearm) and scalp (e.g., motor regions on the left hemisphere) is used to classify hand movements. EEG data is further filtered to the frequency bands typically corresponding to motor imagery or motor execution–the alpha and beta bands from 8-12 Hz and 12-30 Hz, respectively.

Initially, manually labeled segments such as ‘reach’, ‘grasp’, ‘release’, and ‘retract’ were used for leave-one-out session-independent classification in a supervised manner. To classify these actions using EEG and EMG data, we use a combined CNN/LSTM approach as described in [51]. This analysis will provide the basis for additional custom ELAN recognizer plugin development, to generate activity annotations based on multimodal biosignals.

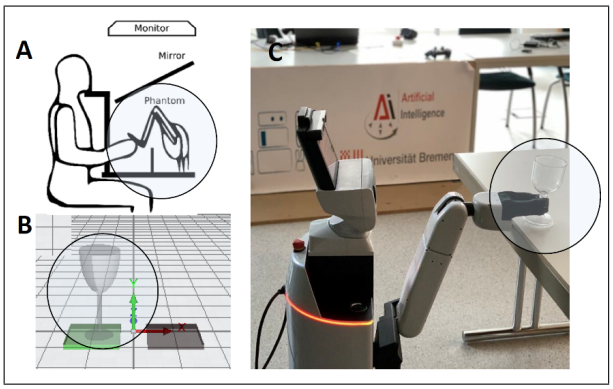

Fig. 6: Transfer of human data to robot control. (A) Data assessment using the PHANToM haptic interface is (B) combined with VR presentation, to (C) control a robot.

VI. Iterative Execution and Improvement through Integration with ELAN and openEASE

The collection of data ultimately serves the purpose of improving the performance of robot systems and enabling them to execute certain actions within the given context and environment. To make the collected data available to the openEASE robotics platform, the detected activities and objects must be translated into a format known to the robot–high-level action plans, such as `move to position Z` or `place object X on surface Y`.

As depicted in figure 6, a pilot demonstration has shown that human data from the EASE-MAD can be successfully transferred to robot actions. In this use case, the task was to place a delicate object, such as a fragile wine glass, on a table. The data were assessed in VR using the PHANToM haptic device in order to present the subjects with realistically rendered forces during placing actions along with the visual sensory feedback. The idea was to transfer the skill of a fast movement with force control to the robot. The resulting end effector variables were suitable to enable the robot to perform the action in an appropriate fashion, i.e. in a real world application it would have been able to lift and to place the glass without breaking it. This approach only comprises a limited range of variables for a short sequence of actions. More complex plans, even though rather abstract in nature, can be executed by the CRAM framework [52].

VII. Results

A. Human Activity Recognition with Multimodal CNNs

We could show that for activity recognition an early fusion approach is better suited than the classic late fusion [48]. For the evaluation in this work, we have used an early fusion architecture with RGB video and optical flow as modalities. We have evaluated the performance on a cross-subject split of the EASE Table Setting Dataset where one recording session (session 17) featuring a specific subject was used as the validation set while seven other sessions featuring other subjects were used as training set. The time slice was set to $0.53s$ (16 frames with 30 fps) for both training and evaluation. With this setting there 29872 data items for training and 8107 for validation. We have achieved an accuracy of 87.8% for the multi-label task and 80.6% for the single-label task on the validation set. Figure 7 shows a plot from one of the training sessions.

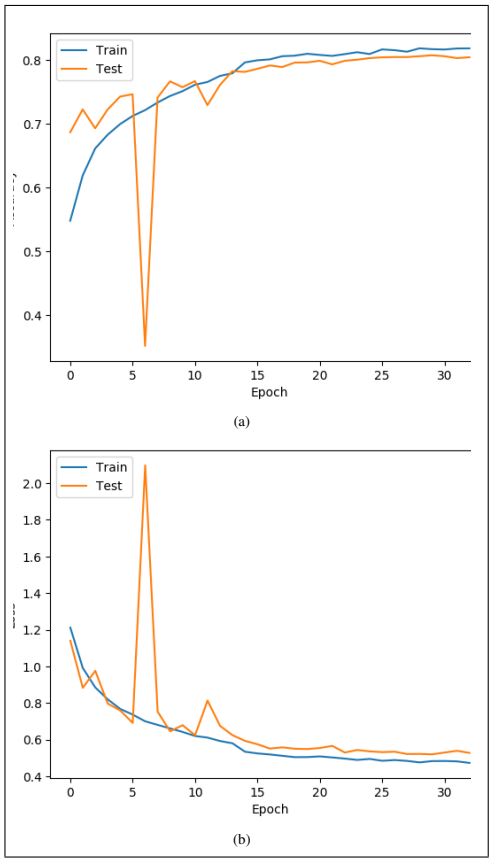

Fig. 7: (a) Accuracy and (b) loss plot for training (blue) and validation (orange) set.

B. Multimodal Biosignal Action Recognition

Brain activity of 30 participants was measured in an EEG-study, consisting of four 1st-person Videos. The videos were annotated according to EASE-ontology, resulting in 312 distinct episodes of various categories and complexity levels. An fMRI-study with 30 participants consisting of ten 1st-person videos was recently finished, with an overall number of 1461 annotated episodes. Preliminary ICA results from this study point out brain areas that discriminate between object interaction events and episodes of no object interaction during the presentation of the videos, as illustrated exemplarily in figure 8. These will later serve as seed regions for the analysis of EEG data.

For the person-independent multi-class motion classification of EASE-TSD trials using a convolutional and long-short term memory (CLSTM) model on EEG and EMG data, the results indicate that EMG features alone provided a better basis for classification at this level of activity. While the low level segments were too brief to extract meaningful information from the EEG sensor data, classification performance

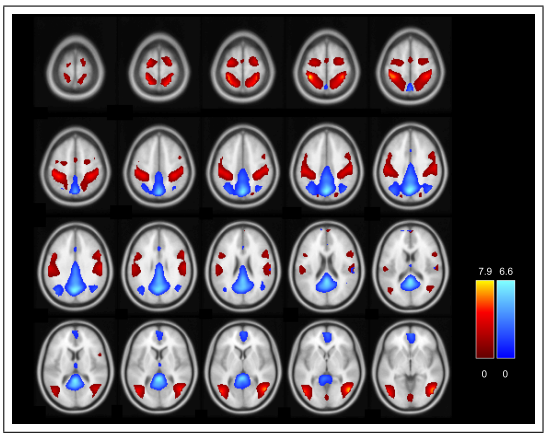

Fig. 8: Brain areas susceptible to stimuli of object interaction events (red) and events with no discernible interaction (blue) during the presentation of a table setting video.

on features derived from EMG sensor data reached 59% accuracy for the right hand movements (MR) and 61.3% accuracy for the left hand movements (ML). Precision for ML was 0.97 vs 1.00 for MR features, recall for ML was 0.95 vs 0.92 for MR, and f1-scores for ML were 0.97 vs 0.95 for MR. Confusion matrices for all combined ML and MR runs are shown in figure 9.

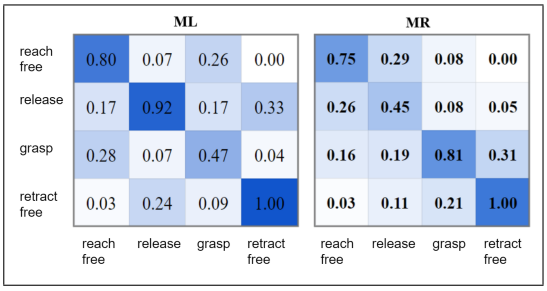

Fig. 9: Confusion matrices for classifications of motions performed with the left and right hand for 4 classes.

VIII. CONCLUSION

Through large-scale collection of human activities of daily living data, annotation with contextually relevant and ontologically linked labeling schema, analysis with diverse multimodal methods for a wide range of sensor modalities, and ultimately, incorporation into the OpenEASE robotics cloud platform, the EASE human activities data analysis pipeline provides the rich groundwork on which to build cognitively-enhanced robotics for use in everyday scenarios.

ACKNOWLEDGMENT

The research reported in this paper has been supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 "EASE - Everyday Activity Science and Engineering", University of Bremen (http://www.ease-crc.org/). The research was conducted in sub-projects H01 *Acquiring activity models by situating people in virtual environments* and H03 *Descriptive models of human everyday activity*.

REFERENCES

[1] M. Beetz, M. Tenorth, and J. Winkler, "Open-EASE – a knowledge processing service for robots and robotics/ai researchers," in *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, Washington, USA, 2015, finalist for the Best Cognitive Robotics Paper Award.

[2] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *CoRR*, vol. abs/1711.08362, 2017. [Online]. Available: http://arxiv.org/abs/1711.08362

[3] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The EPIC-KITCHENS dataset," *CoRR*, vol. abs/1804.02748, 2018. [Online]. Available: http://arxiv.org/abs/1804.02748

[4] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ser. UbiComp '13. New York, NY, USA: ACM, 2013, pp. 729–738. [Online]. Available: http://doi.acm.org/10.1145/2493432.2493482

[5] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," 06 2012, pp. 1194–1201.

[6] M. Tenorth, J. Bandouch, and M. Beetz, "The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1089–1096.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4489–4497. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.510

[8] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 140–153.

[9] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1725–1732. [Online]. Available: https://doi.org/10.1109/CVPR.2014.223

[11] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 568–576. [Online]. Available: http://dl.acm.org/citation.cfm?id=2968826.2968890

[12] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4724–4733.

[13] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[14] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[15] Y. Shi, Y. Tian, Y. Wang, W. Zeng, and T. Huang, "Learning long-term dependencies for action recognition with a biologically-inspired deep network," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[16] D.-A. Huang, L. Fei-Fei, and J. C. Niebles, "Connectionist temporal modeling for weakly supervised action labeling," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 137–153.

[17] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles, "Sst: Single-stream temporal action proposals," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2599174

[19] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, vol. 46, pp. 498–509, 2016.

[21] H. Rahmani and A. Mian, "3d action recognition from novel viewpoints," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 1506–1515. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.167

[22] Z. Shi and T.-K. Kim, "Learning and refining of privileged information-based rnns for action recognition from depth sequences," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] T. S. Kim and A. Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 1623–1631.

[24] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Nov 2015, pp. 579–583.

[25] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807–811, March 2018.

[26] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, May 2017.

[27] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, 2015.

[28] Y. Du, Y. Fu, and L. Wang, "Representation learning of temporal dynamics for skeleton-based action recognition," *IEEE Transactions on Image Processing*, vol. 25, pp. 3010–3022, 2016.

[29] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 4041–4049. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.460

[30] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "Learning with hierarchical-deep models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1958–1971, 2013. [Online]. Available: https://doi.org/10.1109/TPAMI.2012.269

[31] D. Wu and L. Shao, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 724–731. [Online]. Available: https://doi.org/10.1109/CVPR.2014.98

[32] K. Ramirez-Amaro, M. Beetz, and G. Cheng, "Transferring skills to humanoid robots by extracting semantic representations from observations of human activities," *Artificial Intelligence*, vol. 247, pp. 95 – 118, 2017, special Issue on AI and Robotics. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370215001320

[33] D. Triboan, L. Chen, F. Chen, and Z. Wang, "A semantics-based approach to sensor data segmentation in real-time activity recognition," *Future Generation Computer Systems*, vol. 93, pp. 224 – 236, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X18303947

[34] C. Mason, M. Meier, F. Ahrens, T. Fehr, M. Herrmann, F. Putze, and T. Schultz, "Human activities data collection and labeling using a think-aloud protocol in a table setting scenario," in *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges, Madrid, Spain*, 2018.

[35] M. Meier, C. Mason, F. Putze, and T. Schultz, "Comparative analysis of think-aloud methods for everyday activities in the context of cognitive robotics," *20th Annual Conference of the International Speech Communication Association*, vol. 9, p. 10, 2019.

[36] K. A. Ericsson and A. S. Herbert, "Verbal reports as data," vol. 87, no. 3, pp. 215–251, 1980.

[37] M. Jeannerod, "Mental imagery in the motor context," *Neuropsychologia*, vol. 33, no. 11, pp. 1419–1432, nov 1995. [Online]. Available: https://doi.org/10.1016/0028-3932(95)00073-C

[38] R. Der, G. Martius, and R. Pfeifer, *The Playful Machine: Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Springer Science & Business Media, 2012, vol. 15.

[39] C. Zetzsche, J. Wolter, C. Galbraith, and K. Schill, "Representation of space: Image-like or sensorimotor?" *Spatial Vision*, vol. 22, no. 5, pp. 409–424, 2009.

[40] T. Kluss, N. Schult, K. Schill, C. Zetzsche, and M. Fahle, "Spatial alignment of the senses: The role of audition in eye-hand-coordination," *i-Perception*, vol. 2, no. 8, pp. 939–939, 2011.

[41] T. H. Massie, J. K. Salisbury, *et al.*, "The phantom haptic interface: A device for probing virtual objects," in *Proceedings of the ASME winter annual meeting, symposium on haptic interfaces for virtual environment and teleoperator systems*, vol. 55, no. 1. Chicago, IL, 1994, pp. 295–300.

[42] M. C. Çavuşoğlu, D. Feygin, and F. Tendick, "A critical study of the mechanical and electrical properties of the phantom haptic interface and improvements for highperformance control," *Presence: Teleoperators & Virtual Environments*, vol. 11, no. 6, pp. 555–568, 2002.

[43] J. L. Maldonado Cañon, T. Kluss, and C. Zetzsche, "Adaptivity of end effector motor control under different sensory conditions: Experiments with humans in virtual reality and robotic applications," *Frontiers in Robotics and AI*, vol. 6, p. 63, 2019.

[44] J. L. Maldonado Cañon, T. Kluss, and C. Zetzsche, "Pre-contact kinematic features for the categorization of contact events as intended or unintended," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020, pp. 764–765.

[45] J. L. Maldonado Cañon, T. Kluss, and C. Zetzsche, "Categorization of contact events as intended or unintended using pre-contact kinematic features," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 2020, pp. 46–51.

[46] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, Feb. 2015. [Online]. Available: https://doi.org/10.1145/2682899

[47] C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: Association for Computing Machinery, 2005, p. 399–402. [Online]. Available: https://doi.org/10.1145/1101149.1101236

[48] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proceedings of the 23rd International Conference on Information Fusion*. IEEE, July 2020.

[49] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and et al., "Tensorflow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, ser. OSDI'16. USA: USENIX Association, 2016, p. 265–283.

[50] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2017/IMSKDB17

[51] A. Kondinska, "Comparison and classification of multimodal biosignals during and prior to hand motor executions involved in activities of daily living," M.S. thesis, University of Bremen, Germany, 2020.

[52] M. Beetz, L. Mösenlechner, and M. Tenorth, "CRAM – A Cognitive Robot Abstract Machine for Everyday Manipulation in Human Environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, October 18-22 2010, pp. 1012–1017.