Recognition of Bimanual Manipulation Categories in RGB-D Human Demonstration

Franziska Krebs*, Leonie Leven*, Tamim Asfour

***Abstract*— Humans exhibit outstanding capabilities in using both hands to perform daily tasks. Understanding bimanuality in human demonstrations is key for humanoid robots, which should learn from human observation. In this paper, we address the problem of the recognition and segmentation of bimanual action categories defined by our Bimanual Manipulation Taxonomy, based on RGB-D data. To this end, we combine object detection and human motion tracking methods to derive graph-based representations of bimanual manipulation tasks that describe spatial relations between objects and hands as well as the temporal change of these relations during the execution of the task. We train a Graph Neural Network (GNN) for simultaneous recognition and segmentation of the demonstrations and compare the results with a rule-based classification approach that only takes contact relations between objects and hands into account. For training, five kitchen tasks of the KIT Bimanual Actions Dataset are selected and complemented with two new tasks accounting for symmetrical bimanual manipulations. The evaluations show the best results for a GNN considering spatial relations and object knowledge compared to a GNN considering only contact relations between objects and hands and compared to the rule-based approach.**

I. Introduction

A promising approach for teaching robots new skills is Learning from Demonstration (LfD) [1], [2]. The goal is to provide non-expert users with an intuitive way to program robots simply by demonstrating the task. In this paper, we contribute to the programming of bimanual manipulation tasks from human demonstration. In [3], we proposed the *Bimanual Manipulation Taxonomy* that defines different categories of bimanual manipulation actions. These bimanual categories are defined based on different aspects of bimanuality such as the coordination and physical interaction between both hands, the role of each hand in the task, and the symmetry of arm movements during the execution of the task. A detailed description of the bimanual categories defined by the taxonomy is given in Section II-A. To analyze our taxonomy, we applied a rule-based classification of bimanual categories using the KIT Bimanual Manipulation Dataset [3], which consists of bimanual human motion recordings obtained by a marker-based capture system. In this paper, we extend our work by a novel approach for the recognition and classification of bimanual categories – which simultaneously provides a segmentation of bimanual human demonstrations – based on RGB-D data. We consider these bimanual categories crucial for representing bimanual manipulation tasks and the selection of and switching between controllers for the different phases of a bimanual task. The knowledge of the current category can indicate which aspects of the demonstration are crucial (e.g. relative pose) and therefore assist in extracting the relevant task constraints. Being able to recognize these categories in RGB-D data is an important requirement for the usability of the taxonomy in real-world applications on robots equipped with RGB-D sensors.

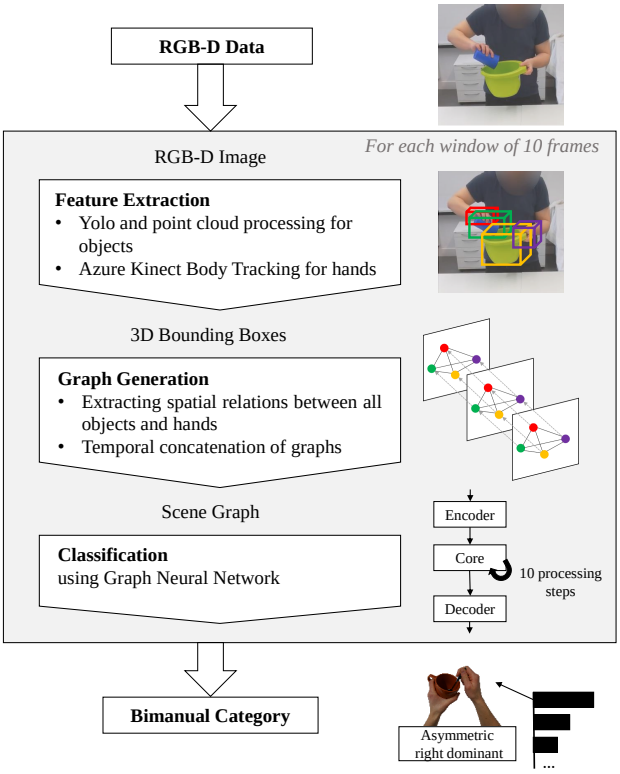


Figure 1. Recognition of bimanual categories in RGB-D data.

Contribution: *(i)* We introduce an approach for the recognition and classification of bimanual categories in RGB-D data based on Graph Neural Networks (GNNs). *(ii)* We extend the KIT Bimanual Actions Dataset (Bimacs) [4] with two additional tasks, focusing on tightly-coupled bimanual symmetric tasks. *(iii)* We compare our previous rule-based classification of bimanual categories with the novel GNN-based approach and demonstrate the benefit of using object knowledge and relations between objects for improved recog-

* The first two authors contributed equally to this work.

This work has been supported by the German Federal Ministry of Education and Research (BMBF) under the project CATCH-HEMI (01KU2012) and by the Carl Zeiss Foundation under the JuBot project.

The authors are with the Institute for Anthropomatics and Robotics, High Performance Humanoid Technologies Lab (H2T), at the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany. {franziska.krebs, asfour}@kit.edu and leonie.leven@student.kit.edu

nition of bimanual categories. An overview of our approach is given in Figure 1.

II. Related Work

This work builds up on the concept of bimanual categories defined by the Bimanual Manipulation Taxonomy in [3]. Thus, we will describe first the categories as well as the previously developed method for their detection. In addition, we will discuss related approaches addressing action recognition with focus on bimanual manipulation tasks.

A. Bimanual Action Categories

In [3], we presented a taxonomy for bimanual manipulation, which is particularly designed for application in robotics. The key aspects which are considered are coordination and interaction between the hands, the roles of each hand in a bimanual task, and the symmetry of the arm movement during the execution of the task. The taxonomy is depicted in Figure 2. On the highest level, the categories are divided into coordinated and uncoordinated. Unimanual left and right and uncoordinated bimanual actions constitute the uncoordinated part of the taxonomy. The branch of coordinated bimanual actions is subdivided based on whether there is physical interaction between the hands or not (tightly vs. loosely coupled). Within the tightly coupled category, we differentiate between 1) symmetric actions, in which both hands have similar roles, and 2) asymmetric actions, in which one hand exhibits a dominant behavior (right- or left-dominant). A rule-based classification was suggested and applied frame-wise on marker-based motion capture data. Thereby, the categories are determined based on contact relations between hands and objects as well as motion features. By doing so, we demonstrated that the categories suggested by the taxonomy can be detected based on high-confidence motion capture data and accurate object models. However, the availability of such data cannot be assumed for real-world robot applications.

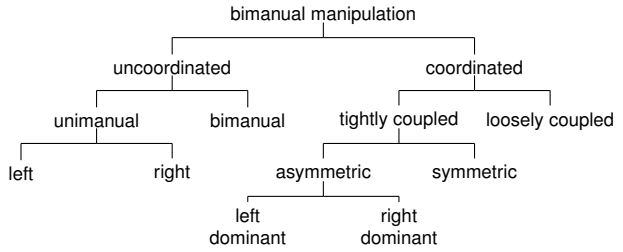


Figure 2. Bimanual Manipulation Taxonomy presented in [3].

Other works address the question of the recognition and classification of bimanual manipulation actions. Boehm et al. [5] use a rule-based classification for the recognition of bimanual coordination modes in the context of robot-assisted surgery. These modes are defined by the direction (e. g., move together or away) and a symmetry (e. g., point or mirror) of the movements. To improve bimanual interaction with a semi-autonomously controlled prosthetic hand, Volkmar et al. [6] distinguish between unimanual, bimanual asynchronous and bimanual synchronous movements. Those are detected by a rule-based classification of the data of two inertial measurement units (IMUs) attached to the prosthesis and the other hand. Miller et al. [7] recognize similar categories based on motion symmetry for the purpose of monitoring the rehabilitation of post-stroke patients. Artificial neural networks are applied on raw IMU data and features extracted from the raw data. An approach for teleoperation of dual-arm robots was proposed by Rakita et al. [8] based on a so-called *bimanual action vocabulary* consisting of "one hand seeking", "self-handover", "fixed offset" and "one hand fixed". A sequence-to-sequence recurrent neural network is used to infer the most probable bimanual action category. The existing approaches are difficult to compare quantitatively since defined categories are different and the analysis was performed based on different sensory modalities. Several approaches ([3], [5], [6]) use rule-based methods which are intuitive and directly allow for error-analysis in case of failed classification. Other approaches use neural networks motivated by handling noisy and chaotic data [7] or inspired by neural processes in humans [8].

In contrast to other approaches, in this work we consider the recognition and classification of seven different bimanual categories as defined by our bimanual manipulation taxonomy [3] and given in Table I. In addition, we aim at the recognition of these categories based on RGB-D data.

TABLE I
Bimanual action categories.

Bimanual Category	Abbreviation
No action	*no_action*
Unimanual left	*uni_left*
Unimanual right	*uni_right*
Loosely coupled & uncoordinated bimanual	*loosely*
Tightly coupled asymmetrical left dominant	*tightly_asym_left*
Tightly coupled asymmetrical right dominant	*tightly_asym_right*
Tightly coupled symmetrical	*tightly_sym*

B. Action Recognition

As discussed in Section II-A there are only a few works addressing the classification of bimanual manipulation action categories and – to our best knowledge – no approaches to solving the problem based on RDB-D or RGB sensory data. However, the problem can also be more generally seen as Human Activity Recognition (HAL) problem, since the main goal is to identify a semantic label to perform human activities based on time-series data obtained from different sensors. Therefore, we will discuss also related work in the area of human action recognition. While action recognition can be directly performed on RGB and/or depth data, we focus on approaches that use previously extracted features. We hypothesize that in a comprehensive LfD framework, the tracking of the human body and the involved objects is essential since this provides the trajectory-level information which is also required. We argue that approaches based on these previously extracted features generalize better to novel

tasks and environments since they are independent of variations in the background and variations in body appearance. Khare and Kumar [9] recently presented a survey on deep learning and RGB-D based human action, human-human and human-object interaction recognition. They provide an overview of relevant datasets and the most commonly used techniques. For human-object interaction detection, which deals with the detection of manipulation-related actions, the survey reveals that most used approaches rely on graph-based representations. Another recent review for human action recognition focuses on the usage of different sensor modalities [10]. Aside from RGB and additional depth data, they name skeleton-based action recognition as a separate modality. The approaches applied for this modality can be divided into three categories: Recurrent Neural Networks (RNNs) including their gaited variants such as Long Short-Term Memories (LSTMs) ([11], [12]), Convolutional Neural Networks (CNNs) ([13], [14]) and Graph Neural Networks (GNNs) ([15], [4]). Notably, GNNs do not only preserve the expressive power of graph structures but are also suitable to handle different input sizes.

While the skeleton modality, as considered in [10], only describes the human body in manipulation tasks, the objects involved are also relevant. This is what is explicitly considered in [4], where a GNN is used for action recognition and segmentation while considering both hands individually. The underlying graph-based representation consists of 1) nodes, which correspond to the hands and the objects detected in the scene, and 2) edges connecting the nodes and describing the spatial relations between hands and objects. Apart from considering spatial relations instead of only contacts to construct a scene graph, the input information used is very similar to the rule-based classification presented in [3] for the detection of bimanual action categories. This suggests that a similar approach might be successful for the recognition of bimanual categories.

III. Approach

The goal of the work is to develop an approach for the recognition of bimanual manipulation categories in household tasks based on RGB-D data and learning-based approaches (GNNs). In this section, we describe the different steps of our approach, as shown in Figure 1, including the creation of suitable training data, the extraction of features, the generation of graph-based representations and the classification using GNNs.

A. Dataset

Learning-based methods such as GNNs require sufficient training data in which all classes of bimanual categories are adequately covered. Thus, we need a dataset that fulfills the following criteria: *(i)* The dataset includes natural human demonstrations of household tasks with bimanual actions, and *(ii)* it resembles data as it would be available for a robot, namely single-view RGB-D data from an observer position. Both criteria are fulfilled for the Bimacs dataset [4], which provides RGB-D data of 6 subjects performing 9 different actions. Each task was recorded 10 times using a PrimeSense Carmine 1.09 camera. The dataset additionally provides extracted 3D bounding boxes of objects and spatial relations of those. Since neither object nor human pose tracking are a focus of this paper, we directly use the data provided. However, we exclude workshop-like assembly tasks and use only the tasks *Cooking*, *Cooking with bowls*, *Pouring*, *Wiping* and *Cereals*. After manually labeling the data with bimanual categories, a closer analysis revealed that the tightly-coupled symmetrical category is hardly included (< 1 % of frames). Therefore, as an extension of the Bimacs dataset, two new tasks were recorded.

1) Data collection: Based on the original dataset, six subjects were recorded (3 male, 3 female; 5 right-handed, 1 left-handed) using an RGB-D camera (Azure Kinect DK). This study was approved by the ethics committee of the Karlsruhe Institute of Technology, Karlsruhe, Germany. The participants gave their written informed consent before the experiments that the data may be made publicly available for research purposes. In the collected video data faces are anonymized. Data was collected at 30 fps with a resolution of 1920px $\times$ 1080px for RGB and 640px $\times$ 576px for depth data. The structure was designed to match the data of the Bimacs database [4]. Two new household tasks were each recorded 10 times for each subject. Subjects were provided with a description of the overall goal to be achieved, but the precise execution was left to them. Initial object positions were varied within different recordings. The two new tasks are designed to include symmetric actions within a household context, namely: *set table* and *prepare dough*. In total six different objects are used: *cup*, *bowl*, *rolling pin*, *spoon*, *whisk* and *plate*. Symmetric motions are expected for transferring big objects like a bowl or a plate and for using a rolling pin. The two new tasks are shown in Figure 3. In total, we collected 120 new recordings with a total duration of 60.2 minutes. Together with the selected recordings of the Bimacs dataset, this resulted in a new dataset with 420 recordings and a total duration of 127 minutes, which we refer to as the combined dataset in the remainder of the paper.

2) Feature Extraction: We use YOLOv7 [16] (trained on the objects in our dataset) to detect the object in RGB video data. The resulting 2D bounding box is used to extract relevant points from a point cloud derived from the depth image. Those were filtered based on the color properties of the detected object and then a 3D bounding box with the dimensions of the object is placed in the center of the remaining point cloud. For tracking the hands, the Azure Kinect Body Tracking SDK was used. A 3D bounding box was constructed based on the minimum and maximum coordinates of the four detected key points per hand (wrist, hand, tip of the hand and thumb).

As a last step we use the methods provided in [17] to extract 15 static and dynamic spatial relations between the extracted 3D bounding boxes as defined in [18], namely *contact*, *above*, *below*, *left of*, *right of*, *in front of*, *behind of*, *inside*, *surround*, *moving together*, *halting together*, *fixed moving together*, *getting close*, *moving apart*, *stable*. The

Figure 3. Exemplary frames from the two recorded tasks: *Set table* (lower row) and *Prepare dough* (upper row).

resulting 3D bounding boxes are transformed to a local coordinate system defined by the ArUco markers placed in the corners of the table, before computing their spatial relations.

3) Labels: Both the data from the Bimacs dataset and the new data were manually annotated in accordance with the definitions of the bimanual categories as defined by the bimanual manipulation taxonomy in [3]. Those labels are used for training the GNN and regarded as ground truth data for the evaluation. Thanks to the extension of the Bimacs dataset, the proportion of tightly coupled symmetric actions in the resulting combined dataset is increased to 13.55 %.

4) Data Augmentation: Due to the uneven representation of right and left-handed subjects, categories such as *uni_left* and *tightly_asym_left* are underrepresented. To compensate for this and ensure better comparability between subjects, we double the data by mirroring it. Therefore, the bounding boxes for the right and left hand are switched and the spatial relations are adapted accordingly by swapping the relations *right of* and *left of*. Furthermore, the labels are also adapted by switching between corresponding left and right categories. This essentially leads to having a right-handed and left-handed version for each subject.

B. Graph Neural Network

For the significantly lower data accuracy extracted from RGB-D data compared to motion capture data, a decrease in the performance of the rule-based approach is expected. Therefore, we consider alternative methods used in action recognition. As stated in Section II deep neural networks are commonly used for this purpose. In our case, special requirements must be fulfilled. On the one hand, the network must be capable of dealing with with variable input sizes due to the variable structure of the considered scene resulting from different numbers of objects and spatial relations. On the other hand, the network should be able to process graph-based representations similar to the scene graphs used in our previous rule-based approach. Graph Neural Networks (GNNs) are predestined for such tasks as they fulfill both requirements. Thus, GNNs are selected for the recognition of bimanual manipulation categories.

As described in [4] and originally defined in [19], we define a graph G as a 3-tuple $G = (u, V, E)$, with u being the global attribute of the graph, V the set of nodes in the graph and E the set of edges. The set of nodes V consists of the node attributes $v_a \in V$ and the edges E of 3-tuples $e = (e_a, s, r) \in E$. Within the edges, e_a represents the edges attributes and s and r are the sender and receiver nodes in V. In our case, the input graph is constructed based on the extracted features in each frame, where nodes are the object/hand instances and edges encode the spatial relations between them. Furthermore, the scene graph of the current and the last nine frames are concatenated by temporal edges connecting the node of a specific object instance between consecutive frames. The global attribute u is not used in the input graph but encodes the determined category as one-hot-encoding in the output graph.

Due to the similarity of the problem in [4] having essentially the same input graphs, a similar network structure is used. The model is formed as in [4] by two independent graph network blocks for the encoder and decoder, respectively, and one full graph network block for the core. The *encode-process-decode* configuration is used. For all blocks multilayer perceptrons (MLPs) were employed as update functions. The sum function was used as aggregation function. All MLPs in each graph network block were parameterized with 2 layers and 256 neurons per layer. The core model performed 10 processing steps. These parameters were empirically determined after evaluating multiple test series on our data.

IV. Evaluation

In this section, we evaluate our approach by comparing the results obtained using GNNs with those of the rule-based approach described in our previous work [3]. To ensure consistency between the Bimacs dataset and its extension, the evaluation is performed separately on the Bimacs dataset, its extension as well as the new combined dataset. The combined dataset consists of 420 recordings with a total duration of about 127 minutes. 53.6 % of the frames belong to the Bimacs and 47.4 % are part of the new recordings. Finally, an ablation study is conducted, to evaluate the influence of features such as temporal graph concatenation and the consideration of spatial relations between hand/object instances instead of only considering contact relations as in the rule-based approach.

A. Graph Neural Network

For the following evaluations, we trained the GNN on the extended Bimacs dataset, which consists of the household tasks of the Bimacs dataset and the new additionally recorded tasks (*set table* and *prepare dough*). As described

in Section III-A.4, the data was doubled for each subject to account for differences in right- and left-handed subjects. To account for the overrepresentation of *loosely* coupled bimanual actions with over 40 % of all frames for the training of the GNN, two of three frames labeled as *loosely* are skipped. This results in a distribution of the ground truth data for the bimanual categories as shown in Figure 4.

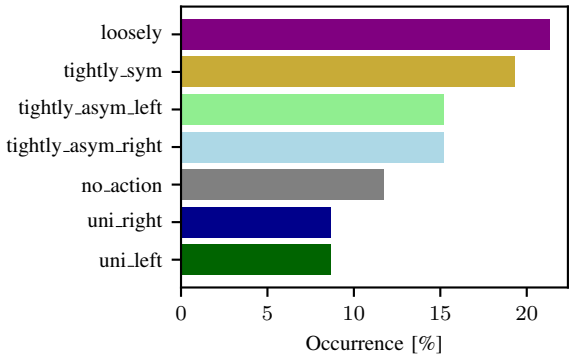


Figure 4. Distribution of the ground truth data.

Following the work of [4] we used the encoder-core-decoder architecture with MLPs as update functions. To obtain the best fitting parameters of these MLPs and the network structure, we performed a systematic evaluation starting with the initial parameters used in [4]. As can be seen in Table II, one of the highest F_1 scores is obtained from the highlighted row, which are our final parameters. The parameters with slightly higher F_1 scores due to a larger history size or a larger amount of processing steps were discarded because of the significantly longer training time.

For training, the dataset was split into a training set and a testing set. Testing sets contain all recordings from one subject (one subject of each dataset including its mirrored motions), while training sets contain all remaining recordings. Additionally, before training, one out of the ten repetitions for each task in the training set was put aside as a validation set. For the quantitative evaluation of the classifier, a leave-one-subject-out cross-validation was performed to obtain six folds of training and testing sets. A combined evaluation of the six test sets results in a macro F_1 score of 0.70. The confusion matrix is shown in Figure 5. The overfitting of the *loosely* category is visible and is caused by the fact that the *loosely* category is the category with the highest occurrence in the training data. However, we hypothesize that this distribution is legitimate for the training data, as the dataset suggests that the category *loosely* is indeed more prevalent in natural movements than other categories.

Table III shows the macro metrics obtained in the evaluation for the different categories. It can be observed that particularly the tightly-coupled categories are detected best. The precision and F_1 score are also high for *loosely*. However, the recall is significantly lower, and as shown in the confusion matrix in Figure 5 many unimanual motions are falsely detected as *loosely*. In the case described above

TABLE II

PARAMETER EVALUATION OF THE MLPS. THE EVALUATION IS ONLY PERFORMED ON ONE SUBJECT. THE HIGHLIGHTED ROW PRESENTS OUR FINAL PARAMETERS.

Layers	Neurons	Batch size	Learning rate	History size	Process steps	Macro F_1-score
2	256	256	0.001	10	10	0.7086
1	256	256	0.001	10	10	0.6870
3	256	256	0.001	10	10	0.6996
2	**128**	256	0.001	10	10	0.6892
2	**512**	256	0.001	10	10	0.7045
2	256	**32**	0.001	10	10	0.6874
2	256	**128**	0.001	10	10	0.6851
2	256	**512**	0.001	10	10	0.6980
2	256	256	**0.01**	10	10	0.5029
2	256	256	**0.0001**	10	10	0.6734
2	256	256	0.001	**1**	10	0.6463
2	256	256	0.001	**5**	10	0.6886
2	256	256	0.001	**20**	10	0.7027
2	256	256	0.001	10	**5**	0.6882
2	256	256	0.001	10	**20**	0.7101

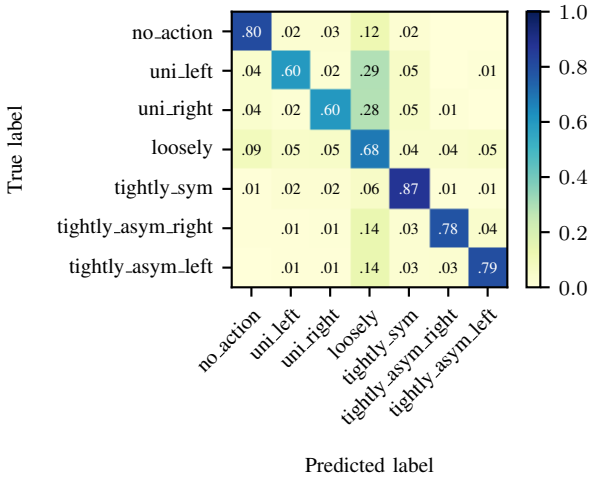


Figure 5. Normalized confusion matrix using GNN with object knowledge.

the node IDs of the input graphs correspond to a specific hand or object for the entire dataset. The GNN does not necessarily know the semantic properties of an object, e.g., a rolling pin, but it does know that when it is used, e.g., when rolling, the symmetric category is often recognized. Therefore, in the case of inference, it can use the object ID for classification, meaning that it has of object knowledge. To avoid this, we kept the object IDs the same only within one recording. This would correspond to the case where the object is unknown, but it can be tracked through a demonstration. For different recordings, the IDs are assigned differently, so that it is not possible for the model to learn a relation as described above. This results in a lower F_1 score of 0.54 and the confusion matrix depicted in Figure 6. As can be seen from the confusion matrix, the results are worse across all categories showing the relevance and advantage of

TABLE III

METRICS OF THE GNN-BASED APPROACH WITH OBJECT KNOWLEDGE.

Category	Precision	Recall	F_1-score
tightly_sym	0.79	0.87	0.83
tightly_asym_right	0.77	0.78	0.78
tightly_asym_left	0.75	0.79	0.77
loosely	0.79	0.68	0.73
no_action	0.59	0.80	0.67
uni_left	0.56	0.60	0.58
uni_right	0.54	0.60	0.57

using object knowledge.

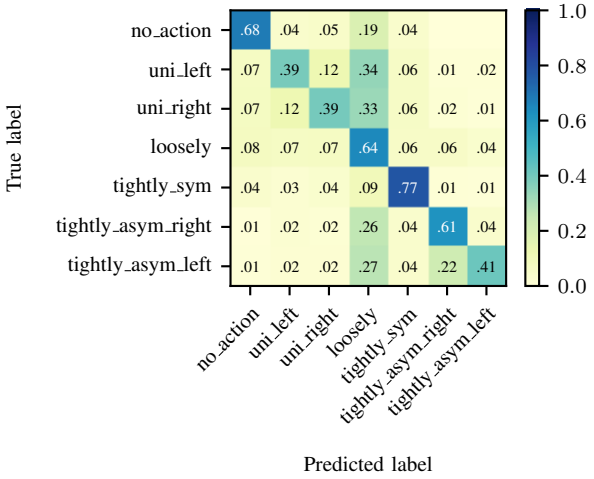


Figure 6. Normalized confusion matrix using GNN without object knowledge.

B. Comparison with Rule-Based Approach

In order to assess the performance of the GNN-based approach, we use the rule-based classification approach described in [3] as a baseline. This approach uses graph-based representations with objects and hands as graph nodes and contact relations between them as graph edges. Therefore, minor adaptions are needed: The consideration of the orientation is discarded since axis-aligned bounding boxes are used. Furthermore, heuristics were implemented to handle objects not detected in intermediate frames and objects only detected in few isolated frames. Threshold parameters are adapted to better suit the less precise data. Figure 7 shows the confusion matrix for the rule-based classification based on the combined dataset including both the original Bimacs dataset and the extension.

The results for the rule-based approach (see Figure 7) are significantly worse than the corresponding GNN-based approach without object knowledge (see Figure 6). This is also evident by the macro F_1 score which improved from 0.40 to 0.54. Predictions are particularly imprecise for the tightly coupled categories. This indicates that for most cases, the contact-based differentiation still works reasonably

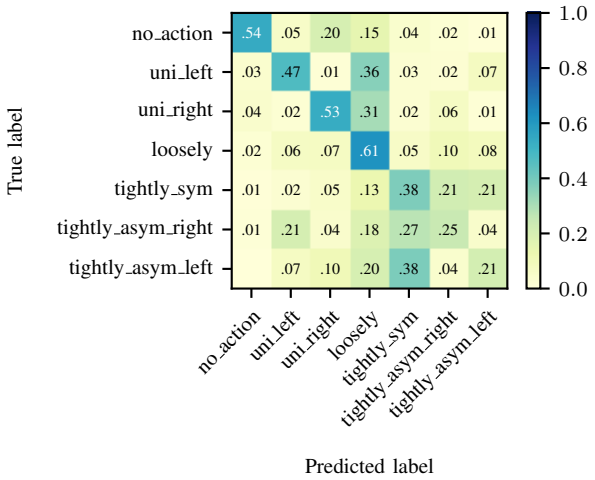


Figure 7. Normalized confusion matrix using the rule-based approach.

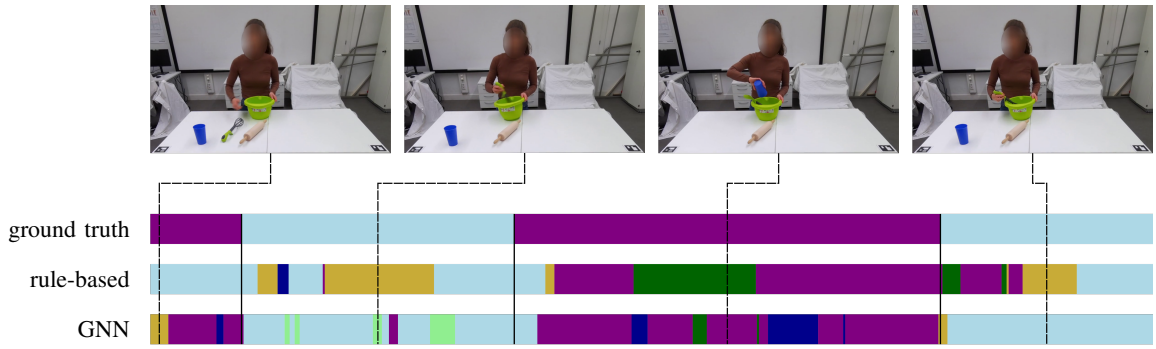
well, but the motion-based differentiation within the tightly coupled categories is not working anymore. This is due to the fact that a rule-based approach with fixed thresholds performs worse with the information extracted from the noisy RGB-D data. In addition, not considering orientations due to using axis-aligned bounding boxes might increase the effect. The high number of frames that are wrongly classified as *loosely* supports this hypothesis. There are some categories whose true labels belong to the tightly coupled categories indicating that a misclassification occurred due to failure in contact detection. However, there are even more frames with true labels that are either *unimanual* or *no_action*. This is also either due to wrong classification, missing objects or imprecise motion data.

When compared to the results of the rule-based approach with motion capture data as presented in [3], the results are – as expected – considerably worse.This is due to the reduced quality of RGB-D data compared to accurate motion capture data used in [3], which makes the detection of objects and hands more difficult and thus leads to wrong classification.

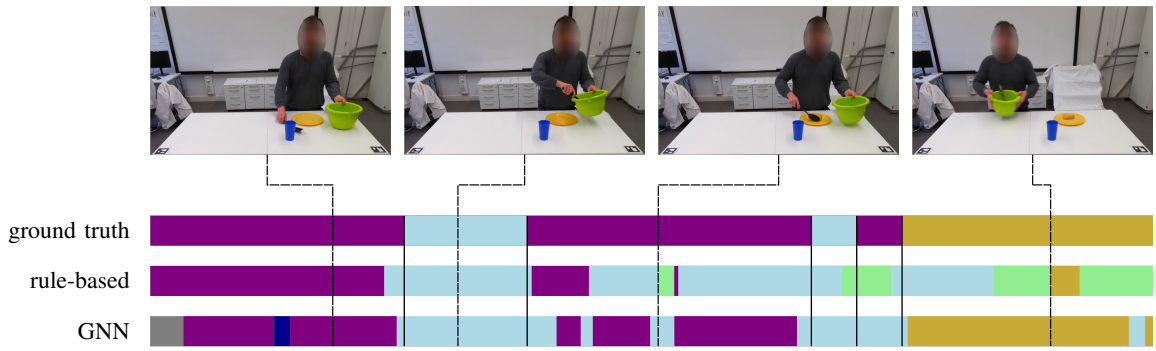
C. Performance on Different Datasets

To ensure a certain level of consistency between the Bimacs dataset and the data recorded in the scope of this work, we compare the classification results. The results of the rule-based approach on the Bimacs, the new data and the combined dataset is shown in Table IV. Parameters are optimized for each dataset individually. The results for the Bimacs and the new recordings are in a similar range, however, the results for the new recordings are slightly better. This is probably due to a more precise tracking of hands and objects.

Since the datasets on their own do not provide enough data and category coverage to train a GNN, we only compare the results of the rule-based approach.



(a) Example segmentation of the *Prepare dough* task.



(b) Example segmentation of the *Set table* task.

Figure 8. Example segmentation of the *Prepare dough* and *Set table* tasks. The top bar visualizes the ground truth, the middle bar the segmentation of the rule-based approach, the bottom bar the segmentation of the GNN-based approach. Categories: *no_action*, *uni_right*, *uni_left*, *loosely*, *tightly_sym*, *tightly_asym_right*, *tightly_asym_left*

TABLE IV

MACRO METRICS FOR THE RULE-BASED APPROACH ON DIFFERENT DATASETS.

Training data	Precision	Recall	F_1-score
Bimacs	0.39	0.45	0.38
New recordings	0.46	0.52	0.46
Combined	0.41	0.43	0.40

D. Ablation Study

As stated in Section III-B, a Graph Neural Network performs significantly better than a rule-based approach for imperfect data and features extracted from RGB-D. In this section, we analyze the specific features of the GNN, which can be adapted for best results. The GNN uses spatial relations instead of only contact relations as done in the rule-based approach in [3]. Therefore, we also train a GNN by only considering contact relations. Furthermore, we consider a version where the input graph contains information from one frame only and there are no temporal edges connecting object instances in the scene graphs over multiple frames. This is evaluated both with spatial relations and only contact relations. We also add the version without object knowledge for comparison. The results are shown in Table V. The resulting metrics of training the GNN without object knowledge (as described in Section IV-A) are also added to the table as a comparison.

As expected the suggested approach performs best. However, interestingly not considering the temporal edges decreases the performance less than considering only contact relations. This could be due to the fact that some temporal information are encoded in the dynamic spatial relations e. g., *halting together*, *moving apart*. As expected the lowest scores are obtained for the model considering only contacts and no temporal edges.

TABLE V

ABLATION STUDY COMPARING THE MACRO METRICS.

Training data			Results		
Spatial Relations*	Temporal Edges	Object Knowledge	Precision	Recall	F_1-score
x	x	x	0.68	0.73	0.70
-	x	x	0.58	0.65	0.60
x	-	x	0.65	0.69	0.65
-	-	x	0.51	0.57	0.52
x	x	-	0.54	0.56	0.54

*In case of no spatial relations only contact relations are considered.

E. Segmentation Results

While the previous section mainly considered the classification, in this section we consider the segmentation. An

exemplary segmentation for an extract of the newly recorded task *Prepare dough* is shown in Figure 8(a) and for the task *Set table* in Figure 8(b). The manually annotated ground truth segmentation is compared against the rule-based and GNN-based approach. Compared to the rule-based approach the segmentation points of the GNN are quite close to the ground truth data. During the *loosely* actions in Figure 8(a) there are some segments of unimanual actions in both approaches which means, that the activity of one hand was not properly detected. For the rule-based approach, the *tightly_asym_right* actions also have a high misclassification rate, on the one hand, because of the threshold for symmetric motions (wrong label *tightly_sym*), on the other hand, because of not recognized contact relations (wrong label *loosely*). This is also evident in Figure 8(b) where particularly the rule-based approach is erroneous within the *loosely* segment and hardly detects the *symmetric* at all.

V. Conclusion

We presented a simultaneous segmentation and classification into the bimanual categories as suggested in [3] based on RGB-D data. To this end, we extended the Bimacs dataset with two additional tasks to cover also symmetrical tasks. We showed that employing a Graph Neural Network (GNN) yields clearly improved results compared to a rule-based approach, especially when object knowledge is provided, and even more so when spatial relations are taken into account instead of just contact relations. There are several aspects that should be discussed considering the presented evaluation. First of all, it has to be mentioned that the manually labeled ground truth data is imprecise, especially considering the exact segmentation points. Nevertheless, we consider it sufficiently well for this purpose. When employing either the GNN-based or the rule-based approach for new tasks there will be several factors influencing the quality of the outcome: Firstly, the rule-based approach is unsupervised and can thus be applied out of the box. It is to be evaluated how well the GNN generalizes for completely different data. Secondly, both approaches highly depend on the quality of the methods for object and body tracking which are employed and can profit from advances in computer vision.

In the future, we will also evaluate the performance of more sophisticated recent GNN architectures as e.g. [20], [21] for the segmentation and recognition of bimanual categories. We plan to combine the segmentation based on bimanual categories with an action segmentation as both provide relevant information for learning comprehensive task models from human demonstration and would benefit from one another. Further, we will work on integrating the developed method on a humanoid robot, enabling the robot to detect bimanual categories in real-time.

References

[1] A. G. Billard, S. Calinon, and R. Dillmann, "Learning from humans," *Springer handbook of robotics*, pp. 1995–2014, 2016.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[3] F. Krebs and T. Asfour, "A bimanual manipulation taxonomy," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11031–11038, 2022.

[4] C. R. Dreher, M. Wächter, and T. Asfour, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 187–194, 2019.

[5] J. R. Boehm, N. P. Fey, and A. M. Fey, "Online recognition of bimanual coordination provides important context for movement data in bimanual teleoperated robots," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6248–6255, 2021.

[6] R. Volkmar, S. Dosen, J. Gonzalez-Vargas, M. Baum, and M. Markovic, "Improving bimanual interaction with a prosthesis using semi-autonomous control," *Journal of NeuroEngineering and Rehabilitation*, vol. 16, p. 140, 2019.

[7] A. Miller and E. Wade, "Classifying unimanual and bimanual upper extremity tasks in individuals post-stroke," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 6301–6305, 2021.

[8] D. Rakita, B. Mutlu, M. Gleicher, and L. M. Hiatt, "Shared control–based bimanual robot manipulation," *Science Robotics*, vol. 4, 2019.

[9] P. Khaire and P. Kumar, "Deep learning and rgb-d based human action, human–human and human–object interaction recognition: A survey," *Journal of Visual Communication and Image Representation*, vol. 86, p. 103531, 2022.

[10] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[11] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention lstm networks for 3D action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1647–1656, 2017.

[12] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, 2017.

[13] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the ACM Multimedia Asia*, pp. 1–6, 2019.

[14] Y. Tang, X. Liu, X. Yu, D. Zhang, J. Lu, and J. Zhou, "Learning from temporal spatial cubism for cross-dataset skeleton-based action recognition," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1–24, 2022.

[15] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7912–7921, 2019.

[16] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[17] R. Kartmann, F. Paus, M. Grotz, and T. Asfour, "Extraction of physically plausible support relations to predict and validate manipulation action effects," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 4, pp. 3991–3998, 2018.

[18] F. Ziaeetabar, T. Kulvicius, M. Tamosiunaite, and F. Wörgötter, "Recognition and prediction of manipulation actions using enriched semantic event chains," *Robotics and Autonomous Systems*, vol. 110, pp. 173–188, 2018.

[19] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[20] H. Xing and D. Burschka, "Understanding spatio-temporal relations in human-object interaction using pyramid graph convolutional network," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5195–5201, IEEE, 2022.

[21] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16041–16050, 2021.