Incorporation of the Intended Task into a Vision-based Grasp Type Predictor for Multi-fingered Robotic Grasping

Niko Kleer[1], Ole Keil[2], Martin Feick[1], Amr Gomaa[1], Tim Schwartz[1], and Michael Feld[1]

***Abstract*— Robots that make use of multi-fingered or fully anthropomorphic end-effectors can engage in highly complex manipulation tasks. However, the choice of a suitable grasp for manipulating an object is strongly influenced by factors such as the physical properties of an object and the intended task. This makes predicting an appropriate grasping pose for carrying out a concrete task notably challenging. At the same time, current grasp type predictors rarely consider the task as a part of the prediction process. This work proposes a learning model that considers the task in addition to an object's visual features for predicting a suitable grasp type. Furthermore, we generate a synthetic dataset by simulating robotic grasps on 3D object models based on the BarrettHand end-effector. With an angular similarity of 0.9 and above, our model achieves competitive prediction results compared to grasp type predictors that do not consider the intended task for learning grasps. Finally, to foster research in the field, we make our synthesized dataset available to the research community.**

I. INTRODUCTION

Robotic grasping plays a significant role in numerous fields such as manufacturing environments, ambient assisted living, or prosthesis applications. Traditionally, robots located in static environments are pre-programmed to repeatedly carry out the exact same task. To this end, they often utilize a parallel jaw gripper and are able to surpass humans in performance and precision [1]. On the other hand, prosthetic hands are usually anthropomorphic, and robots that need to adapt to strongly differing object geometries often utilize multi-fingered end-effectors. While such end-effectors enable more dexterously manipulating objects, planning a suitable grasp increases in complexity. As their design is inspired by the human hand, research from the field of human grasp analysis provides significant insights about the factors influencing the choice of a suitable grasp [2], [3], [4], [5], [6], [7]. It is particularly notable that not only the physical properties of an object influence the choice of a suitable grasp, but the intended tasks associated with an object as well (see Figure 1). Harnessing these insights, numerous approaches for predicting a suitable grasp have emerged.

Grasp type prediction models generally aim to determine a suitable grasp based on object-related features. These methods are examined with a particular focus on the learning methodology [8], [9], [10], explored in the context of prosthesis control [11], [12], [13], [14], [15], [16], [17], and actively utilized in multi-fingered robotic grasping applications [18], [19], [20], [21], [22], [23], [24], [25], [26]. Most approaches utilize computer vision techniques to predict a suitable grasp based on the exclusive use of visual features. In other cases, textual resources, electromyography (EMG) signals, or a mixture of multiple modalities is considered [11], [17]. However, most grasp type predictors do not currently include the intended task in the process of predicting a suitable grasp. For example, whether a pen is supposed to be picked from a flat surface, held for writing, or used during a handover is disregarded. Although there exists awareness about the significance of the task [10], [16], [19], [21], [27], authors often do not consider the aspect or leave its incorporation for future work.

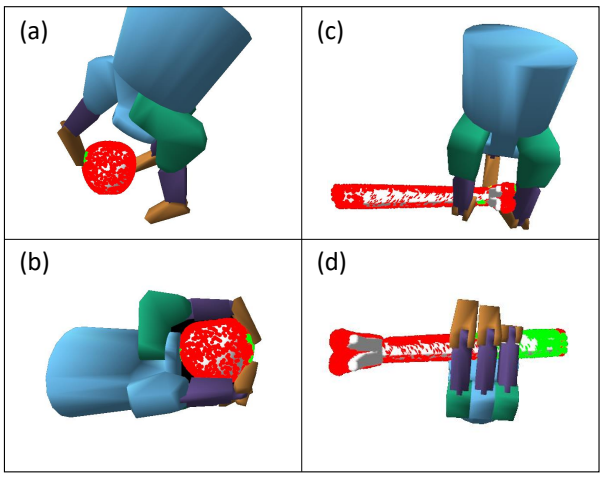

Fig. 1. Task-specific grasps simulated on the BarrettHand end-effector and 3D models of an apple and a hammer. The first column shows the tasks (a) *handoff* and (b) *use* carried out with an apple, while (c) and (d) show grasps for the same tasks applied to the hammer respectively.

While building on top of related research, this work specifically investigates the challenge of incorporating the intended task into vision-based grasp prediction models. This paper makes the following contributions:

- We demonstrate two variations of incorporating the intended task into a vision-based grasp type predictor and show that our method achieves competitive prediction results in comparison to models that do not consider the task for learning grasp types.
- We propose a hierarchical schema for systematically labeling grasp types executable by the BarretHand[1]

[1]DFKI, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany {niko.kleer;martin.feick;amr.gomaa;tim.schwartz;michael.feld}@dfki.de

[2]Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany

[1]https://advanced.barrett.com/barretthand

robotic end-effector.
- We generate a synthetic dataset by simulating multi-fingered robotic grasps on 3D object models and augment the resulting 693 valid (object,task) tuples with information about the executed grasp types. To support research in this field, we make this dataset available to the research community.

II. RELATED WORK

As this research utilizes findings from the field of human grasp analysis and grasp type prediction, we subsequently dedicate a section to the related work of each field.

A. Analysis of Human Grasps

The field of human grasp analysis has experienced extensive developments since Schlesinger [28] categorized grasps into six basic grasp types and Napier [6] first distinguished between the precision and power-grip categories. Based on these categories, Cutkosky [2] later established a more refined taxonomy of human grasps in the form of a hierarchical layout. In addition to considering the physical properties of an object, the taxonomy also includes aspects related to the object manipulation task and constraints related to the environment. More than two decades after its introduction, Feix et al. [3] presented an even more sophisticated categorization of human grasp, which the authors named the GRASP taxonomy. Their work is built on top of the state-of-the-art while including extensive observations on the relation of an object's physical properties and several tasks on the choice of a grasp [4], [5], [29]. It further represents the most complete categorization of human grasps currently established. As the GRASP taxonomy involved the use of quantitative data, Stival et al. [7] investigated the similarity of grasps using qualitative data. They use electromyography (EMG) signals and kinematic data to construct a hierarchical structuring of grasps. Finally, Arapi et al. [30] recently introduced a novel taxonomy of human grasps for video labeling. Their work contributes to the challenge of gathering datasets for data-driven grasp synthesis methods.

The literature presented in this section serves as a basis for grasp type prediction models. We provide an overview of such models in the next section.

B. Grasp Type Predictors

Grasp type predictors exploit the aforementioned knowledge about human grasping in order to predict a canonic grasp type for manipulating an object. They have been explored in numerous contexts and involve the use of various data sources. The majority of these models rely on computer vision-based learning methods. This includes grasp type predictors investigated to control a prosthetic hand [11], [12], [13], [14], [16], [17] and robots that utilize multi-fingered end-effectors [18], [19], [20], [21], [22], [23], [24], [25], [26]. A few approaches have also explored the potential of textual descriptions [9], [24]. In these cases, textual data describing attributes such as the general geometry, size, and material of an object can effectively contribute towards grasp predictions. Furthermore, modeling of grasp types can benefit robotic grasp planning while outperforming similar grasp planning algorithms that do not consider this information [21]. For studying grasp type predictors based on the BarretHand end-effector, Lin and Sun [31] specifically provide a mapping that matches the grasp types introduced by Cutkosky [2] to executable BarretHand grasps. Notably, most grasp type predictors strongly focus on learning object-related features. As a result, the significance of the task is not considered. Specifically, Lu and Hermans [21], who demonstrate the use of grasp types within their probabilistic grasp planning framework, suggest including information about the task for further improving grasp predictions. Moreover, Cai et al. [32] discuss the importance of studying the relation between an object's physical attributes in conjunction with the tasks associated with it. This is further emphasized by Yang et al. [10] who outline how the applied grasp type can have implications on how humans perceive and interpret the executed action. The authors illustrate how the use of different grasp types applied to a knife can convey either a threatening gesture or a simple handover. This can be considered a human-centered aspect that motivates the use of grasp types. Finally, some authors aim to incorporate the intended task in their future work [16], [18], [19].

As only a few works have considered task-related information for grasp type prediction [9], [17], [24], our research investigates the challenge of predicting a suitable grasp based on the visual features of an object and the intended task associated with the grasp. We build on top of the learning architecture proposed by Zandigohar et al. [16] who have previously considered predicting grasp types from visual data and EMG signals [17]. In contrast to our work, their approach is dedicated to controlling a prosthetic hand and has not considered object manipulation tasks. Additionally, our work proposes a hierarchical schema for systematically labeling BarretHand grasps, extending the grasp type mapping introduced by Lin and Sun [31].

III. PROPOSED APPROACH

Our method for integrating task-specific information into a grasp type predictor involves the following steps:

1) We first generate a dataset that combines information about objects, associated object manipulation tasks, and observations about suitable grasp types for carrying out each task. This is particularly important as datasets used in the literature do not combine all these aspects [3], [16], [18].
2) Second, we establish a learning architecture that incorporates the intended task as a feature for learning suitable grasp types.

In the next three sections, we elaborate on how we generate data from synthesized robotic grasps, systematically label our samples, and suggest suitable learning architectures.

A. Grasp Synthesis

One of the main challenges in generating a dataset lies in acquiring information about the grasp types that are suitable

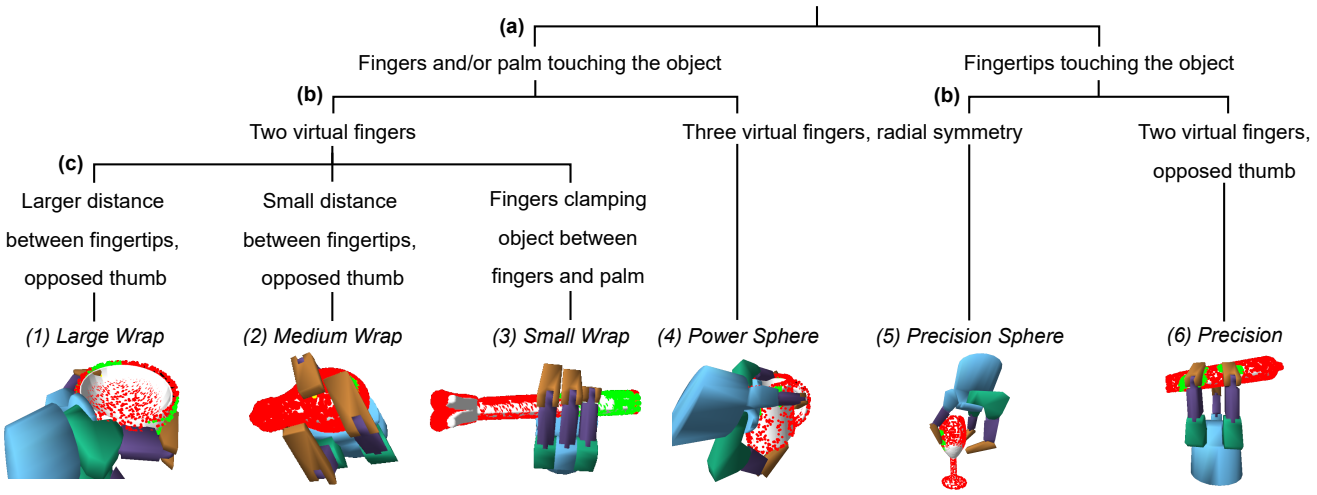

Fig. 2. Hierarchical rule set for an iterative sample labeling procedure. Level (a) distinguishes between more powerful and more precise grasps, (b) separates poses based on the number of virtual fingers involved, and (c) specifically differentiates between different types of "wrap" grasps.

for carrying out a specific task. For example, it seems clear that writing with a pen commonly requires a different grasp as compared to handing it over to somebody else. To learn about the suitability of grasp types for certain tasks, we utilize the ContactGrasp framework introduced by Brahmbatt et al. [33], [34]. The framework allows synthesizing functional grasps on 3D object models based on contact maps for multiple end-effectors using the GraspIt! simulator [35]. A functional grasp is a grasp dedicated to carrying out a specific task. Leveraging the contact maps provided by ContactDB [33], we simulate a total of 20 grasps for the tasks *use* and *handoff* respectively. It is important to note that the task *use* corresponds to a precise action in relation to each object (e.g., talking on the phone or screwing a light bulb into a socket). The grasps are simulated on the BarretHand end-effector. We repeat this procedure for the same 19 objects as the authors of the ContactGrasp framework, resulting in a total of 760 samples. Examples of task-specific synthesized grasps for the objects "apple" and "hammer" are visualized in Figure 1. We used this data as a basis for learning about the grasp types used in specified object manipulation tasks.

While ContactGrasp allows us to synthesize grasps, it does not provide any information about which grasp types are used on an object. Next, we must determine these grasp types for obtaining labeled samples.

B. Sample Labeling Procedure

Before we can train a learning model for predicting a suitable grasp based on an object and a given task, we have to (1) specify the target classes of our model and (2) label each sample accordingly. As we synthesize grasping poses using the BarretHand end-effector, we adopt the grasp type mapping introduced by Lin and Sun [31]. They provide a mapping from human grasps to executable BarretHand grasps based on Cutkosky's taxonomy [2]. However, labeling our data according to these grasps is challenging without a consistent methodology. Therefore, we have developed a schema for systematically labeling each sample. It contains a ruleset for iteratively determining a canonic grasp type applied by the BarretHand end-effector. Our rules can be visualized in a hierarchical structure as demonstrated in Figure 2. First, we distinguish between grasps that involve the use of the fingertips or the entire finger, sometimes in conjunction with the palm. This step resembles the prominent distinction between precision and power grasps. After that, we further separate these grasps based on the number of virtual fingers used during the grasp. The number of virtual fingers describes the directions in which parts of the hand apply force to an object. For example, spherical grasps usually involve a hand configuration where each finger applies force from a different direction. Finally, we distinguish between different "wrap" grasps, which are commonly influenced by the diameter of an object. As a result, we can systematically determine the grasps generated during the grasp synthesis procedure, and label each sample accordingly. The only samples we have not labeled are the ones showing an invalid grasp. Cases of invalid grasps include objects not being

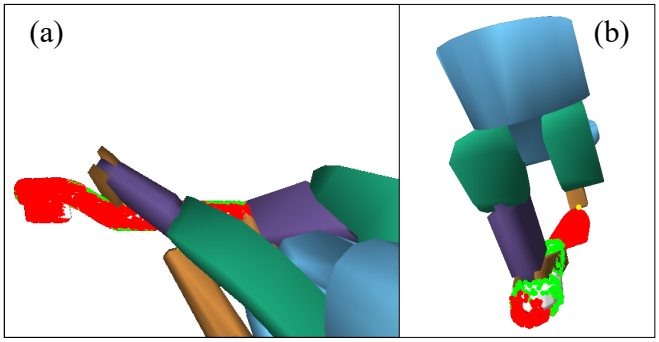

Fig. 3. Examples for invalid grasping poses. The grasp shown in (a) would cause the object to drop and in (b), parts of the gripper penetrate the object.

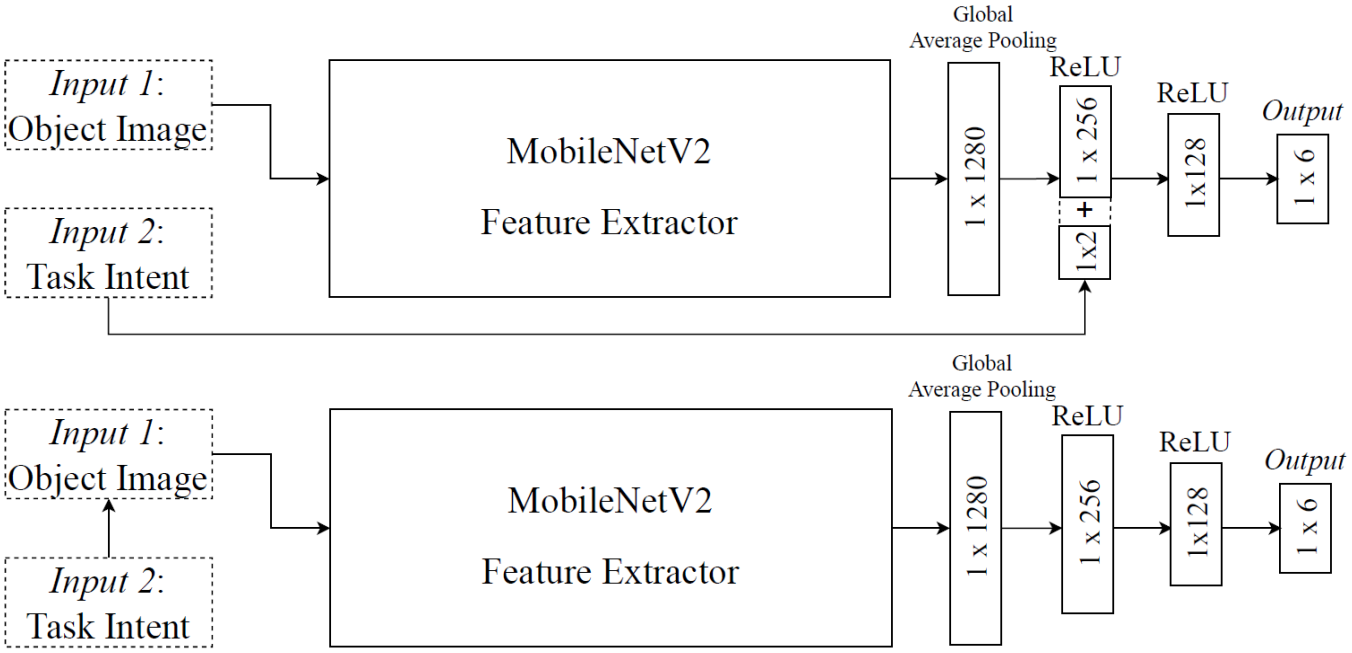

Fig. 4. Proposed learning architectures that integrate the task into the learning process. The upper architecture appends a one-hot encoded vector to the second fully connected layer after feature extraction. The lower architecture directly encodes information about the task into the image of an object.

gripped by the BarrettHand (i.e. the object would drop to the ground due to gravity), or the end-effector penetrates parts of the object. Figure 3 shows an example for either case. Our retrieved data ultimately contains 693 valid samples, combines objects with their associated tasks, and provides a grasp type distribution. The final grasp type distribution corresponds to 52, 59, 144, 197, 103, and 138 grasps in order of their visualization in Figure 2 respectively. We make this dataset[2], which effectively augments the work by Brahmbatt et al. [34] with information about grasp types, publicly available. We also include metadata required for reproducing the synthesized grasps using their framework. For additional comprehensibility, the dataset contains a screenshot of each synthesized grasp. This may enable other authors to create a dataset that suits their research project.

As the final step of our proposed approach, we have to establish a learning model that incorporates the intended task into the process of predicting a suitable grasp.

C. Learning Architectures

Since there has been a strong focus on the use of visual features among researchers developing grasp type prediction models, a large number of convolutional neural network-based model architectures have emerged [8], [10], [12], [14], [16], [18], [19]. However, Zandigohar et al. [16] have compared numerous convolutional neural network architectures and conclude that the MobileNetV2 [36] feature extractor achieves the best results while being most efficient in terms of floating point operations. We choose the MobileNetV2 as the core component of our architecture as this further allows us to compare our results to the approach evaluated by the authors. However, the feature extractor does not account for information about object manipulation tasks. Therefore, we propose the following strategies for integrating the intended task into the learning architecture.

- Our first strategy takes inspiration from Zunjani et al. [37], who distinguish between an unspecified primary and secondary task to determine grasping regions for a robot using a parallel jaw gripper. In contrast to their use case, our model does not predict a region but has to learn a grasp type configuration according to an object and a specified task. To this end, we replace the last layer of the MobileNetV2 feature extractor with four fully-connected layers and append a one-hot encoded vector to one of the layers. During our experiments with the architecture, we determined that appending this vector to the second fully connected layer is the most sensible choice.
- One disadvantage of our first strategy is that the intended task entirely circumvents the feature extractor. As a result, we might not leverage the full potential of the learning model. Therefore, as our second strategy, we propose to directly encode information about the intended task into the image of an object. We explicitly increase and decrease the RBG values of an image for the tasks *use* and *handoff* by the value 10 respectively. As a result, the core component of our model may directly extract information about the intended task.

Both proposed strategies, as well as the exact network architecture, are visualized in Figure 4. In the next section, we elaborate on how we have evaluated our approach based on the presented dataset and learning architectures.

[2]https://github.com/nikleer/GraspTypesWithTasks

IV. EVALUATION

For our evaluation, we implemented the models presented in Section III-C. The MobileNetV2 feature extractor is pre-trained on the well-established ImageNet dataset [38]. For training our final grasp type predictor, we render realistic images of 3D object models similar to the models used by ContactGrasp [34]. Each object is rendered from 100 random viewpoints, resulting in a total of 1900 RGB images with a dimension of 256×256 pixels. Our final training samples combine object images, the intended task, and grasp types for each (object,task) tuple. For the output of our network, we use a binary vector representation. For example, the binary vector $[1, 1, 0, 0, 0, 0]$ indicates that we have observed the grasp types "large wrap" and "medium wrap" at least once, indicating that both grasps are suitable for carrying out the task. To maintain an equal distribution of grasp type labels in our training and test set, we split the data using a stratified 5-fold cross validation. Finally, we use the following evaluation metrics to retrieve the results of our models.

A. Evaluation Metrics

We use three metrics to evaluate our proposed approach. First, we utilize the angular similarity measure, a metric introduced and used by Zandigohar et al. [16], which allows measuring the similarity between two probability distribution vectors. Transforming the distribution of grasp types into a probability distribution is done by dividing each value of a binary vector by the sum of all values (e.g., the binary vector $[1, 1, 0, 0, 0, 0]$ is transformed to $[0.5, 0.5, 0, 0, 0, 0]$). Using this measure enables us to compare our results to their work and observe whether the inclusion of the task leads to a difference in the model's performance. The angular similarity of the input vector u and the predicted vector v is calculated following the formula

$$sim(u, v) = \left(1 - 2 \cdot \arccos\left(\frac{u \cdot v}{||u|| \cdot ||v||}\right) \div \pi\right).$$

We obtain the predicted vector v by applying the Softmax function to the last layer of our learning model. Furthermore, precision, recall, and accuracy represent commonly used evaluation metrics for assessing the performance of machine learning models. Since our method involves a multilabel classification problem, we retrieve the multilabel accuracy

$$accuracy(u, v) = \frac{|u \cap v|}{|u \cup v|}$$

and multilabel precision

$$precision(u, v) = \frac{|u \cap v|}{|v|}$$

to learn more about the overall prediction quality of our models [39]. As these two evaluation metrics require both vectors, u and v, to contain absolute values, we specifically apply the Sigmoid function to our results and consider a grasp type prediction in vector v to be true in case the prediction probability of a class exceeds 0.5. We are unaware of use cases that may ignore false positive predictions, which is why we do not retrieve multilabel recall.

B. Results

Based on the evaluation metrics previously described, we have retrieved results for different configurations of our models. All results are displayed in Table I. Model

TABLE I

ANGULAR SIMILARITY, MULTILABEL ACCURACY, AND MULTILABEL PRECISION OF DIFFERENT MODEL CONFIGURATIONS.

Model	$sim(u, v)$	$acc(u, v)$	$prec(u, v)$
(a) Model without task	0.81	0.85	0.87
(b) Task as vector	0.90	0.88	0.88
(c) Task in image	0.92	0.92	0.92
(d) Model ignores task	0.94	0.98	0.99
(e) Zandigohar et al. [16]	0.93	×	×

variation (a), which achieves the worst performance in all retrieved metrics, represents our learning model without a mechanism to account for the associated tasks. This means that the implemented architecture does not include the one-hot encoded vector and the information about the task was not encoded into the object images. On the other hand, a direct comparison of our proposed learning architectures (b) and (c) to the results retrieved by Zandigohar et al. [16] shows that we achieve competitive angular similarities while considering the task. Encoding the information about the task into the image results in a minor difference of 0.01 in angular similarity. This shows that tasks can be integrated into a vision-based grasp predictor without suffering a major loss in prediction quality. Our retrieved multilabel accuracies and precisions are equivalent to the angular similarity of model (c) and nearly match the angular similarity of model (b), indicating no inconsistencies. Finally, we have also evaluated a configuration of our model that ignores the task. This means that all observed grasp types were considered valid for manipulating an object. For example, the observed distribution vectors $[1, 1, 0, 0, 0, 0]$ and $[0, 0, 0, 1, 0, 0]$ are combined into a joint vector $[1, 1, 0, 1, 0, 0]$ and the model is not required to distinguish between different tasks. It is similar to vision-based grasp type predictors that do not consider the task and the aspect is implicitly assumed during the data labeling procedure. However, the assumption automatically leads to a degree of arbitrariness during grasp choice, which we aim to diminish. In real-world scenarios where a robot has to distinguish between explicit tasks that require different grasping gestures applied to the same object, such assumptions might not be possible. This model's architecture resembles model (e) the most. Due to this simplification, the approach achieves the highest angular similarity, multilabel precision, and multilabel accuracy since the choice of a grasp is independent of the object manipulation task. In the next section, we discuss noteworthy aspects related to our results and elaborate on potential future work.

V. DISCUSSION

Based on our evaluation, it is evident that a learning architecture without a mechanism to factor in object manipulation tasks performs notably worse than the other models. While

our analysis is based on a limited set of objects and tasks, it highlights the impact of the associated object manipulation tasks related to the model's ability to choose a suitable grasp. This can likely be attributed to differing grasp type distributions for the tasks associated with an object, which poses a challenge when trying to learn grasp types from object images alone. An analysis of our data shows that there exists an average of 3.5 valid grasp types for the (object,task) tuples. Upon closer examination of the grasp type distribution, we observe that there is an average of 1.1 grasp types occurring for different tasks involving the same object. We hypothesize that predicting appropriate grasps will become even more challenging as more objects and tasks are added to such a system. This suggests that it is necessary to have a deeper understanding of how to incorporate tasks into a grasp predictor. Even though our proposed approaches provide ways for dealing with this challenge, it is worth discussing that both architectures make potentially limiting assumptions about the system.

Our results show that encoding task-related information into the image of an object may benefit the grasp type predictor. It represents our method with the best performance in all retrieved metrics while considering the task. However, this approach could encounter scalability issues as soon as too many tasks must be distinguished. To overcome this limitation, a more sophisticated method would be required. Alternatively, including an additional image that demonstrates the task being executed could serve as input for the feature extractor. By producing a collection of images that illustrate the task being executed from randomly generated perspectives, similar to the object images we utilized, the predictor's capabilities could be further improved. This method not only addresses the aforementioned limitation in a more refined manner but may also enable the extraction of more intricate information about the object manipulation tasks.

In part, the issue can be mitigated by integrating a one-hot encoded vector that captures task-related information into the learning model. While our experiments indicate that this method performs slightly worse in all retrieved evaluation metrics, it offers unique advantages over relying on a single vision-based classification model. In our implementation, this vector can be interpreted as a surrogate for an additional source of information, such as a prediction model that is intended to classify the object manipulation task. By utilizing such a multi-stage classification model, we may use multimodal data and select the most appropriate modality for determining the task that is supposed to be executed. Recent research has demonstrated that multimodal data outperforms using only one modality for predicting grasp types in controlling a prosthetic hand [17]. While the authors used EMG signals for a small number of hand gestures, other sources of information could be integrated into such models as well. Leveraging this methodology, it would also be possible for a human to specify the task during human-robot interactive situations (e.g., by using language or demonstrating the grasp). This could result in creating further incentives to use grasp types for not only enhancing robotic grasp planning but also fostering an interface for human-robot interaction. In particular, viewing grasp types from a human-centered perspective appears to be underexplored in the existing related work. Finally, Since the execution of most object manipulation tasks also depends on factors such as the environment, considering other contextual attributes could likewise contribute to better grasp predictions.

VI. CONCLUSION

In this research, we have examined the incorporation of the intended object manipulation task into a vision-based grasp type predictor for a multi-fingered robotic end-effector. Our approach involves generating a synthesized dataset and enriching the data with information about the executed grasp types. We have developed a rule-based schema for methodical labeling and proposed two learning architectures that integrate the intended task into the model. Our evaluation demonstrates that we can successfully integrate object manipulation tasks into a vision-based grasp type predictor without compromising its prediction accuracy, competing with models that do not factor in the task. Our results align with existing literature that highlights the task's importance in selecting an appropriate grasp. Nevertheless, predicting grasp types remains a complex task due to numerous factors that influence their selection. Therefore, we are interested in further exploring how additional contextual data, such as multimodal data, can be used to enhance grasp type prediction. Specifically, we intend to investigate how these prediction models can be utilized in robotic grasping applications where the system needs to differentiate between multiple object manipulation tasks. Finally, we hope that our work will assist other researchers in overcoming the existing challenges in this field.

ACKNOWLEDGMENT

This work is supported by the German Federal Ministry of Education and Research as a part of CAMELOT - Continuous Adaptive Machine-Learning of Transfer of Control Situations (grant no. 01IW20008) and FedWell - Life-Long Federated User and Mental Modeling for Health and Well-being (grant no. 01IW23004).

REFERENCES

[1] K. Tai, A.-R. El-Sayed, M. Shahriari, M. Biglarbegian, and S. Mahmud, "State of the art robotic grippers and applications," *Robotics*, vol. 5, no. 2, p. 11, 2016.

[2] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 269–279, 1989.

[3] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[4] T. Feix, I. M. Bullock, and A. M. Dollar, "Analysis of human grasping behavior: Correlating tasks, objects and grasps," *IEEE Transactions on haptics*, vol. 7, no. 4, pp. 430–441, 2014.

[5] ——, "Analysis of human grasping behavior: Object characteristics and grasp type," *IEEE Transactions on haptics*, vol. 7, no. 3, pp. 311–323, 2014.

[6] J. R. Napier, "The prehensile movements of the human hand," *The Journal of bone and joint surgery. British volume*, vol. 38, no. 4, pp. 902–913, 1956.

[7] F. Stival, S. Michieletto, M. Cognolato, E. Pagello, H. Müller, and M. Atzori, "A quantitative taxonomy of human hand grasps," *Journal of neuroengineering and rehabilitation*, vol. 16, no. 1, pp. 1–17, 2019.

[8] A. Das, A. Chattopadhyay, F. Alia, and J. Kumari, "Grasp-Pose Prediction for Hand-Held Objects," in *Emerging Technology in Modelling and Graphics*. Springer, 2020, pp. 191–202.

[9] N. Kleer, M. Feick, and M. Feld, "Leveraging publicly available textual object descriptions for anthropomorphic robotic grasp predictions," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7476–7483.

[10] Y. Yang, C. Fermuller, Y. Li, and Y. Aloimonos, "Grasp type revisited: A modern perspective on a classical feature for vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 400–408.

[11] M. Cognolato, M. Atzori, R. Gassert, and H. Müller, "Improving robotic hand prosthesis control with eye tracking and computer vision: A multimodal approach based on the visuomotor behavior of grasping," *Frontiers in artificial intelligence*, vol. 4, p. 744476, 2022.

[12] J. DeGol, A. Akhtar, B. Manja, and T. Bretl, "Automatic grasp selection using a camera in a hand prosthesis," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, pp. 431–434.

[13] G. Ghazaei, "Deep vision for prosthetic grasp," Ph.D. dissertation, Newcastle University, 2019.

[14] L. T. Taverne, M. Cognolato, T. Bützer, R. Gassert, and O. Hilliges, "Video-based prediction of hand-grasp preshaping with application to prosthesis control," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 4975–4982.

[15] F. Vasile, E. Maiettini, G. Pasquale, A. Florio, N. Boccardo, and L. Natale, "Grasp pre-shape selection by synthetic training: Eye-in-hand shared control on the hannes prosthesis," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 13 112–13 119.

[16] M. Zandigohar, M. Han, D. Erdoğmuş, and G. Schirner, "Towards creating a deployable grasp type probability estimator for a prosthetic hand," in *International Workshop on Design, Modeling, and Evaluation of Cyber Physical Systems, Workshop on Embedded Systems and Cyber-Physical Systems Education*. Springer, 2020, pp. 44–58.

[17] M. Zandigohar, M. Han, M. Sharif, S. Y. Gunay, M. P. Furmanek, M. Yarossi, P. Bonato, C. Onal, T. Padir, D. Erdogmus, *et al.*, "Multimodal fusion of emg and vision for human grasp intent inference in prosthetic hand control," *arXiv preprint arXiv:2104.03893*, 2021.

[18] Z. Deng, G. Gao, S. Frintrop, F. Sun, C. Zhang, and J. Zhang, "Attention based visual analysis for fast grasp planning with a multi-fingered robotic hand," *Frontiers in neurorobotics*, vol. 13, p. 60, 2019.

[19] Z. Deng, B. Fang, B. He, and J. Zhang, "An adaptive planning framework for dexterous robotic grasping with grasp type detection," *Robotics and Autonomous Systems*, vol. 140, p. 103727, 2021.

[20] D. Dimou, J. Santos-Victor, and P. Moreno, "Grasp pose sampling for precision grasp types with multi-fingered robotic hands," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 773–779.

[21] Q. Lu and T. Hermans, "Modeling grasp type improves learning-based grasp planning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 784–791, 2019.

[22] Q. Lu, M. Van der Merwe, B. Sundaralingam, and T. Hermans, "Multifingered grasp planning via inference in deep neural networks: Outperforming sampling by learning differentiable models," *IEEE Robotics & Automation Magazine*, vol. 27, no. 2, pp. 55–65, 2020.

[23] J. Lundell, E. Corona, T. N. Le, F. Verdoja, P. Weinzaepfel, G. Rogez, F. Moreno-Noguer, and V. Kyrki, "Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4495–4501.

[24] A. B. Rao, K. Krishnan, and H. He, "Learning robotic grasping strategy based on natural-language object descriptions," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 882–887.

[25] C. Wang, D. Freer, J. Liu, and G.-Z. Yang, "Vision-based automatic control of a 5-fingered assistive robotic manipulator for activities of daily living," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 627–633.

[26] Y. Zhang, J. Hang, T. Zhu, X. Lin, R. Wu, W. Peng, D. Tian, and Y. Sun, "Functionalgrasp: Learning functional grasp for robots via semantic hand-object representation," *IEEE Robotics and Automation Letters*, 2023.

[27] N. Kleer and M. Feick, "A study on the influence of task dependent anthropomorphic grasping poses for everyday objects," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*. IEEE, 2022, pp. 829–836.

[28] G. Schlesinger, "Der mechanische aufbau der künstlichen glieder," in *Ersatzglieder und Arbeitshilfen*. Springer, 1919, pp. 321–661.

[29] I. M. Bullock, J. Z. Zheng, S. De La Rosa, C. Guertler, and A. M. Dollar, "Grasp frequency and usage in daily household and machine shop tasks," *IEEE Transactions on haptics*, vol. 6, no. 3, pp. 296–308, 2013.

[30] V. Arapi, C. Della Santina, G. Averta, A. Bicchi, and M. Bianchi, "Understanding human manipulation with the environment: a novel taxonomy for video labelling," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6537–6544, 2021.

[31] Y. Lin and Y. Sun, "Grasp mapping using locality preserving projections and knn regression," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1076–1081.

[32] M. Cai, K. M. Kitani, and Y. Sato, "Understanding hand-object manipulation with grasp types and object attributes," in *Robotics: Science and Systems*, vol. 3. Ann Arbor, Michigan;, 2016.

[33] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8709–8719.

[34] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2386–2393.

[35] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[37] F. H. Zunjani, S. Sen, H. Shekhar, A. Powale, D. Godnaik, and G. Nandi, "Intent-based object grasping by a robot using deep learning," in *2018 IEEE 8th International Advance Computing Conference (IACC)*. IEEE, 2018, pp. 246–251.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[39] M. S. Sorower, "A literature survey on algorithms for multi-label learning," *Oregon State University, Corvallis*, vol. 18, no. 1, p. 25, 2010.