Capturing Task-related Information for Text-based Grasp Classification using Fine-tuned Embeddings

Niko Kleer[1] [0000−0003−4288−4724], Leon Weyand[2] [0009−0005−7101−6848], Michael Feld[1] [0000−0001−6755−5287], and Klaus Berberich[2] [0000−0003−3813−9721]

[1] DFKI, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany
{niko.kleer;michael.feld}@dfki.de

[2] Saarland University of Applied Sciences, 66123 Saarbrücken, Germany
{lweyand;klaus.berberich}@htwsaar.de

Abstract. Manipulating objects with a robotic hand or gripper is a challenging task that can be supported by knowledge about the object, such as textual descriptions. Even with such knowledge, there remain numerous possibilities for applying an appropriate grasping gesture. This ambiguity can be reduced by providing information about the intended task, aiding robots in making the choice of a suitable grasp less arbitrary and more robust. This work investigates using word embeddings in the context of grasp classification for multi-fingered robots. Instead of predicting grasping gestures without specifying the intended task, our work combines a description of the properties of an object and task-related information. We demonstrate that a systematically generated dataset and fine-tuned context embeddings can compete with existing models that do not consider object manipulation. Our best model achieves a micro f1 score of 0.774 and macro f1 score of 0.731 while distinguishing between over 40 tasks.

Keywords: Grasp Classification · Text Classification · Word Embeddings · Natural Language Understanding

1 Introduction

Grasp classification in multi-fingered robotics aims to determine a canonic grasping pre-shape suitable for grasping or manipulating an object. Well-established literature about human grasping shows that the choice of a grasp depends on numerous factors [3, 10]. It is important to note that these factors not only include physical attributes such as an object's shape or size but also relate to the tasks it can be used for. Even though many approaches have investigated classifying these grasping gestures, or integrated knowledge about them in their system, information about concrete tasks often plays a subordinate role or is not considered [4, 6, 11, 14, 23]. However, a robotic system should be aware of the tasks a grasping gesture is associated with as this may influence how an object must be handled. Imagine an industrial human-robot collaborative setting where simple machine tools such as a screwdriver or a wrench have to be used. In case of using such a tool, the robot likely requires a powerful grip to apply pressure to a specific region. However, handing over the same tool to a human might require a completely different grasp on the object to guarantee the smooth completion of this task. Especially handling objects or workpieces with a wide range of complex object geometries involved in a series of tasks reinforces the challenge of appropriate grasp choice. In the context of spoken instructions or textual descriptions, this problem equates to capturing semantics about the attributes of an object in conjunction with information about the tasks it can be used for.

Conceptually, there is a resemblance to text classification tasks that aim to extract the meaning of a sentence based on the words it contains e.g., during sentiment analysis [12]. Over the last few decades, the field of Natural Language Processing (NLP) has proposed many approaches that consider many factors for capturing the meaning of words or sentences. Prominent methods that model the (co-)occurrences of words include the Term-Frequency Inverse-Document Frequency (TF-IDF) vector space model, Naïve Bayes, or Latent Semantic Analysis (LSA) [5]. Well-known machine learning models that leverage specific architectures for establishing vector representations of words include Word2Vec [19], Global Vectors (GloVe) [22], and the Bidirectional Encoder Representations from Transformers (BERT) [7]. Although a few works have investigated how to predict suitable grasping gestures from natural language [14, 23], it is still unclear how to adequately capture the semantic information of textual descriptions while distinguishing between a wide range of object manipulation tasks.

To improve our understanding of this challenge, we make the following contribution: This paper investigates the capabilities of word embeddings to capture semantic information about objects and their associated tasks for grasp classification. By appropriately accounting for object manipulation, we aim to make the grasp choice less arbitrary and more in line with real-world situations. To this end, we conducted two data collection studies to determine relevant tasks and collect their grasping behavior based on five grasp types. Furthermore, we systematically generated a dataset of textual descriptions of an object's properties in conjunction with these tasks. Finally, we evaluated our approach using multiple embedding models and applied different strategies for dealing with the complex relationship between the objects, the associated tasks, and their corresponding grasping gestures. We demonstrate that fine-tuned context embeddings can successfully capture these relationships while achieving competitive results compared to existing models not distinguishing between tasks.

2 Related Work

This work is primarily related to two fields. We subsequently elaborate on the concept of grasp types, which is fundamental for grasp classification, and continue by briefly discussing the background and development of word embeddings.

2.1 Grasp Types in Robotics

Grasp classification builds on top of a large body of research dedicated to analyzing how humans handle objects of everyday life [1–3, 8–10]. These findings show that the applied grasping gestures can be categorized according to criteria related to the properties of objects and environmental influences. As a result, several categorization approaches that divide frequently re-occurring gestures into so-called *grasp types* were proposed [3, 10]. A grasp type is a canonic grasping pre-shape that describes the placement of the fingers and palm relative to an object. Although the well-established GRASP taxonomy by Feix et al. [10] contains over 30 grasp types, applications involving the use or classification of grasp types usually distinguish between a much smaller number. In particular, research demonstrates that even the most fundamental distinction between the precision and power grasp categories can benefit robots during grasp planning [18]. As the names of these categories suggest, a precision grasp is used for fine-grained manipulation whereas a power grasp often involves the palm to allow for a stable grip. While the computer vision community has put great effort into understanding how to classify grasp types [4, 6, 11, 18], only a few approaches have investigated the use of natural language [14, 23]. The authors Rao et al. [23] proposed extracting nominal attributes that describe e.g., an object's shape, material, hardness, or texture. They applied regular expressions and sentence chunking to obtain these features from textual descriptions that follow similar formatting. Furthermore, the size of an object was measured and included as numerical values to enable their multi-fingered robot to plan and apply a secure final grasp. Building on top of this work, Kleer et al. [14] extracted object descriptions from public websites such as Wikipedia to not constrain their learning model to a predefined set of features. They did not incorporate a dedicated feature extraction procedure and classified grasp types based on unstructured textual descriptions, allowing for more flexibility. However, in both of these cases, the authors classified grasp types for only one task. To our knowledge, there are no publications that investigate how to distinguish between different tasks based on the semantics of a sentence. Therefore, our work explores how to effectively capture the semantic relationship between an object's properties and the associated tasks for grasp classification.

2.2 Development of Word Embeddings

The modeling and extraction of semantic information has long been a topic of interest in NLP. Early approaches involved modeling (co-)occurrences of words as sparse vector representations, also known as the Bag-of-Words (BoW) model. A prominent method that represents words in this model is LSA [5]. By applying a technique called singular value decomposition (SVD), the matrix of word frequencies is reduced to multiple matrices of lower dimensions. This allows capturing the latent relationships between words. On the other hand, the BoW model suffers from several disadvantages such as assuming words to appear independently and its incapability to model more complex relationships (e.g., polysemous words and contextual meaning). Two decades later, Mikolov et al. [19]

proposed a neural architecture to obtain numerical representations, so-called word vectors or embeddings, for the words of a corpus. This approach is known as Word2Vec and builds on the idea that words appearing in the same semantic context have a similar meaning. The authors distinguish between the Continuous Bag-of-Words (CBOW) and Skip-gram model where a word based on a local context window is predicted and vice versa. In an attempt to further improve the modeling of the semantic relationship of words, Pennington et al. [22] later suggested leveraging global word statistics instead of a local context window, known as Global Vectors (GloVe). These word embeddings have shown impressive results in many NLP tasks [25] and several variations for text classifications were proposed, such as Doc2Vec [16] or fastText [13]. One major disadvantage to these methods is that they cannot adequately capture contextual information when processing a sequence of words as there is no mechanism to account for it. This is also why traditional embeddings can face challenges dealing with polysemous words or idioms [20, 21]. To overcome existing challenges, incorporating contextual information through the BERT architecture was proposed [7]. Unlike traditional embeddings, BERT uses a so-called attention mechanism that recognizes the interdependence of contextual words and learns their importance regardless of their distance from each other. Since BERT uses a bidirectional transformer model, it can effectively account for significant contextual words from both directions. This ultimately allows establishing an embedding for each word depending on its surrounding context. Similar to traditional word embeddings, several variations of BERT were developed e.g., an even more optimized version called RoBERTa [17] and a distilled model known as DistilBERT [24]. Grasp classification shares many similarities with text classification tasks where capturing the relationship of words is significant. Yet, it is currently unclear whether embedding-based methods can recognize and account for attributes that dictate grasping behavior when considering object manipulation.

3 Method

To conduct our investigations, we needed to collect information about the relationship between objects, their associated tasks, and suitable grasping gestures. Since the authors Kleer et al. [14] have publicized the list of objects they used for their grasp classification approach, we based our work on the same set of 100 household objects. However, as they have classified grasp types based on only one task (i.e., holding an object), their labeling is insufficient for us. Therefore, we carried out two consecutive data acquisition studies where we asked humans (1) to note down grasping tasks for the objects and (2) to identify the appropriate grasping behavior. After that, we generated textual descriptions that capture the physical characteristics of these objects and added information about the tasks gathered as a part of our study. In the next two sections, we elaborate on the specifics of our data acquisition studies and the object description generation procedure.

3.1 Data Acquisition Studies

Our first study was intended to collect a set of tasks that each object in our dataset can be used for. Feix et al. [10], who have established the GRASP taxonomy, previously distinguished between two task categories that describe reasons for grasping, namely *holding* and *using* an object. While holding an object describes general manipulations, uses typically depend on its purpose. Since this categorization could cause a high degree of ambiguity during grasp classification, we gathered a set of more concrete tasks based on these categories instead. Following, we describe the setup of our first study.

(1) Grasping Tasks Study: We recruited five participants (two females and three males) between the ages of 21 and 28 with no Computer Science background. Each participant was given a document containing a list of the 100 household objects. Furthermore, the document contained instructions for noting down tasks related to both holding and using task categories. To collect hold tasks, the participants were instructed to note down everyday situations or tasks in which they would typically grasp or hold common household objects. The tasks should not be specific to an individual object but applicable to all. On the other hand, to collect use tasks, the participants were instructed to note down tasks each of the 100 household objects can be used for.

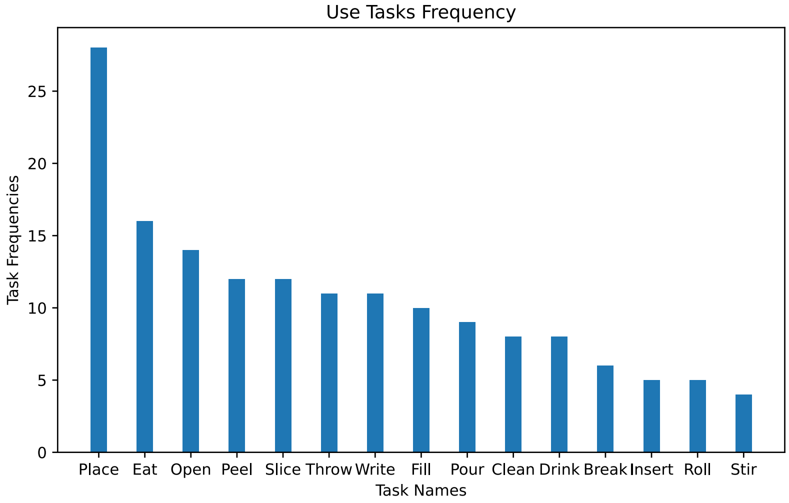

Fig. 1. Most frequently mentioned tasks by our study participants for object use.

To obtain the final set of tasks for each object, we aggregated the answers by our participants and kept all tasks mentioned by an absolute majority (i.e., a task was noted by at least three of them). As a result, our dataset contained over 400 object-task combinations, averaging around four tasks per object. We collected 47 unique tasks, two describing generic hold tasks applicable to all objects, and 45 tasks for specific uses. While the generic tasks correspond to securely gripping an object and handing it over to somebody else, our data covers a wide range of applications an object can be used for (see Figure 1 for the most frequent tasks). We further observe that many tasks are only associated with a few objects. For example, the mobile phone is the only object in the dataset used for making a phone call. However, even though this is the case, it is important to note that completely unrelated tasks might require the same grasping behavior during execution. To learn about the distribution of suitable grasping gestures for the object manipulation tasks we gathered, we conducted a second data acquisition study as described below.

(2) Grasping Behavior Study: For conducting our grasping behavior study, we recruited the same five participants (two females and three males) between the ages of 21 and 28 who participated in our grasping tasks study. This time, each participant was given a document listing all combinations of objects and tasks that resulted from our previous study. Furthermore, the document contained five pictures demonstrating the execution of a grasp type (see Figure 2) and an instruction asking them to assign the most suitable grasp for each combination. During the study, our participants did not know about the names of these grasp types and were only provided the numbers in the upper right corner. The reason for choosing these poses is based on the literature which shows that they are most frequently used to handle objects because of their distinct hand-finger configurations [1, 10].

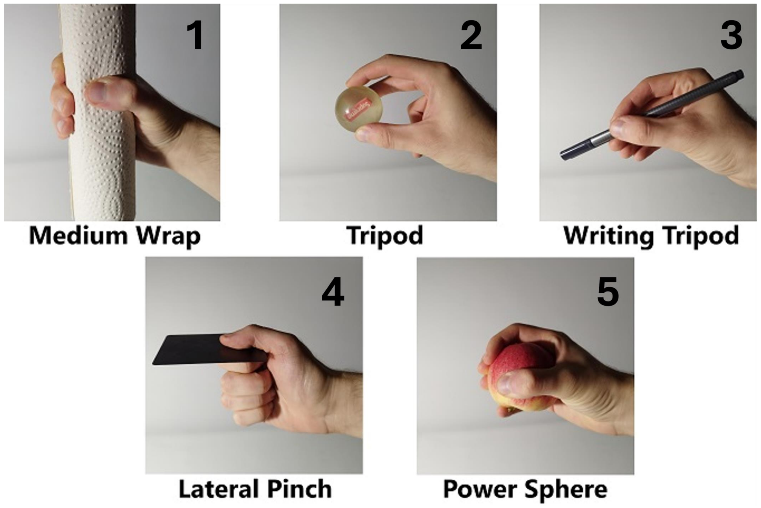

Fig. 2. Five demonstrations of distinct grasp types commonly used by humans for object manipulation according to the literature [1, 10].

Similar to our previous study, we aggregated all answers by our participants and the final label was determined based on a majority vote. This time, reaching an absolute majority was not a requirement as humans might not agree on only one grasping gesture [14]. Our method of label assignment always yielded a unique result, meaning that we did not observe a draw. Our label distribution corresponded to 187 Tripod, 129 Medium Wrap, 71 Lateral Pinch, 19 Power Sphere, and 14 Writing Tripod grasps. This shows a strong imbalance but also highlights the challenge of optimal grasp assignment. We elaborate in our Evaluation Section on how we dealt with these imbalances. Next, we needed to generate textual descriptions that would allow us to capture the attributes of our objects in conjunction with the tasks we gathered.

3.2 Description Generation

The features that influence the choice of a suitable grasp are well-researched [1–3, 8–10]. Based on these insights, we wanted to generate a dataset of descriptions that combines both the attributes of an object and its tasks. To achieve this goal, we leveraged the power of OpenAI's Large Language Model (LLM) ChatGPT 3.5[3]. This approach is similar to newly developed frameworks, such as ROSGPT [15], that provide an interface to ChatGPT for facilitating human-robot interaction. We thoroughly explored strategies for generating descriptions that include useful information for grasp classification and experimented with prompting strategies that influence the description's length, content, and the language used for describing objects in conjunction with their tasks. We faced many challenges and found that the LLM could not adequately generate descriptions of the objects and their tasks at the same time. This is because textual portions describing our tasks were often unclear and focused on the objects rather than the action. For example, instead of describing how to grasp an apple for slicing, the LLM would generate that one has to grab a knife to slice the apple. Therefore, we split this problem into separately generating an object and task description while applying slightly different strategies. In case of the object descriptions, we prompted the LLM to:

- *Imagine the object in front of you*: This restricted the LLM to describing tendencies rather than all theoretically existing variations of an object. Excluding the constraint resulted in too general descriptions of object attributes such as "the object can occur as either heavy or light".
- *Describe in two to three sentences*: When we did not specify the description's length, the LLM generated long paragraphs including a considerable amount of geographic or historical information unrelated to grasping behavior.
- *Mention shape, material, hardness, texture, fragility, and weight*: These attributes were derived by Rao et al. [23] and can tremendously influence the handling of objects. Besides mass, they were defined as discrete sets of words, which is different from our work.

[3] https://chat.openai.com/

- *Exclude quantifiable features*: Dealing with numerical features can be difficult, especially because the LLM usually generates comprehensive paragraphs about the dimensions and weight of an object. We excluded this information from our investigations and focused on the semantics of words.
- *Keep the description objective*: We introduced this constraint to remove evaluative terms that do not influence grasping behavior.

Since we noticed during our investigations that the LLM could not adequately describe and integrate information about our tasks, we chose to generate a separate task description using a different strategy. For each task in our dataset, we first formulated a concise sentence that outlines the action, independently of the LLM. For example, we described the task of securely holding an object as "Hold the object firmly while carrying to prevent accidental drops" and the handover task as "Hold the object to hand it over to another person". We created such a description for each of our 47 unique grasping tasks. After that, we prompted the LLM *to provide a synonym for this description by using different words and grammar while maintaining its meaning.* This resulted in the LLM producing descriptions that focused on the action rather than the object. Additionally, we prompted the LLM to produce an object and task description based on simple, everyday, and formal language to increase the amount of data. Finally, we concatenated the object and task descriptions for each object-task pairing, resulting in 1260 samples. The next section elaborates on how we used this data to evaluate our grasp classification approach.

4 Evaluation

We evaluated our labeled descriptions for the challenge of grasp classification by implementing a series of learning models. Similar to the authors Kleer et al. [14], we used a Naïve Bayes (BoW) classifier, a Support Vector Machine (SVM) based on the TF-IDF vector space model, and a Convolutional Neural Network (CNN) where the initial layer contained pre-trained GloVe [22] embeddings. Our CNN uses the same layering architecture to warrant a fair comparison to their work. Rao et al. [23] used a NN whose inputs were restricted to nominal variables extracted from their descriptions. Their architecture does not apply to our approach and was not implemented. Our evaluation further includes a comparison of pre-trained GloVe and Word2Vec embeddings whose dimensions were 300. Finally, we included fine-tuned DistilBERT context embeddings. In terms of hyperparameters, all networks were trained for 30 epochs using a batch size of 8, the Adam optimizer, and our data followed an 80/20 training-test split. We employed an early stopping condition where the patience hyperparameter was set to five to prevent our models from overfitting on the training data. This means if a model could not improve its performance within the next five epochs after obtaining the best result, it maintains the model weights before these epochs. Our experiments demonstrated that lower values result in stopping the training too early while higher values might not benefit the training as the early stopping condition is never met. We evaluated three model variations to deal with the

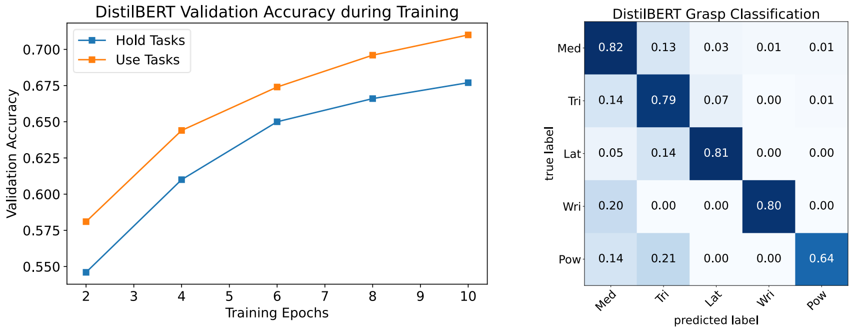

Fig. 3. Validation accuracy progression over the first 10 epochs (left) and final confusion matrix (right) of our best model. All values are averages.

challenge of imbalanced label distribution (see our grasping behavior data collection study in Section 3.1). Specifically (1) a base model (i.e., data and model were unmodified), (2) a model where our dataset was randomly undersampled based on the minority class, and (3) a weighted model where the class weights were chosen inversely proportional to their frequency in the dataset. We did not include the weighting approach for the Naïve Bayes and DistilBERT model as there is no intuitive parameterization option for these weights. Table 1 provides an overview of the micro f1 (accuracy) and macro f1 scores for each model.

Table 1. The best performances of our grasp classification models where the highest average Micro F1 and Macro F1 scores of each model are highlighted.

Model	Base		Undersampling		Weighted	
	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
Naïve Bayes (BoW)	0.584	0.429	0.575	0.566	X	X
SVM (TF-IDF)	0.679	0.489	0.652	0.638	0.690	0.638
CNN (GloVe)	0.729	0.665	0.654	0.631	0.736	0.706
CNN (Word2Vec)	0.722	0.650	0.646	0.621	0.734	0.695
DistilBERT	0.774	0.731	0.699	0.689	X	X

Our results show that the Naïve Bayes and SVM classifiers cannot compete with the embedding-based approaches. This is different compared to the work by Kleer et al. [14], whose best model is an SVM with an accuracy of 0.79. Although we can achieve significant improvements in macro f1 by undersampling the data or adding suitable class weights, the micro f1 score remains under 0.7. The Word2Vec and GloVe embeddings are equally powerful for classifying grasp types while distinguishing between many tasks. Similar to the other models, we can increase macro f1 by accounting for minority classes by using suitable class weights. Nonetheless, the DistilBERT context embeddings outperform all other models with a micro f1 score of 0.774 and a macro f1 score of 0.731. This is a difference of less than 0.02 in accuracy compared to the literature where the authors achieve an accuracy of 0.79 while considering only one task [14, 23]. As shown by Figure 3, our context embeddings can equally adapt to both task categories during training and the final model predicts nearly all grasp types with similar accuracy. The power sphere grasp type causes the most confusion and is mainly responsible for the decrease in macro f1. Overall, our approach yields competitive results while distinguishing between almost 50 tasks and considers minority classes that might be significant for object manipulation.

5 Discussion

A qualitative analysis of our results shows that, while the models slowly adapt to most grasping tasks and capture many details during training, a few tasks pose too much of a challenge. The gripping task, for example, was mentioned by our participants for the tweezers and pliers. Here, the models would need to learn that using tweezers necessitates a precision grasp (e.g., writing tripod) whereas using pliers often requires force and a powerful grasp (e.g., medium wrap). This distinction is difficult as it involves a deep understanding of the relationship between an object's properties, tasks, and the applicable grasp types. Collecting data to model these details is very challenging since grasping tasks are unevenly distributed across different objects. Therefore, factoring in this information can significantly increase the complexity of grasp classification. Additionally, we had assumed that integrating grasping tasks into our grasping behavior study (see Section 3.1) would simplify the optimal grasp assignment. Even though this constraint limits the number of feasible grasps on an object, our study participants did not always agree on the same gesture for a particular object-task pairing. This is because many objects can be grasped in different ways, even when the task is specified. We believe that more research is needed on how to model the existing knowledge about human grasping [3, 8–10] to learn appropriate grasping gestures based on specific object manipulation. To address this, it may be feasible to incorporate pre-defined task constraints into a grasp classification model, drawing inspiration from ontological frameworks that assist robots in learning about their environment.

6 Conclusion

This work investigated using word embeddings in the context of grasp classification for multi-fingered robots while focusing on including task-related information. To this end, we conducted two data acquisition studies for systematically gathering related tasks and suitable grasping behavior for 100 household objects. After that, we generated textual descriptions that combine an object's attributes in conjunction with the tasks gathered from our study. Finally, we evaluated our approach using five learning models while employing strategies for coping with the imbalances in our data. Our results demonstrate that context embeddings show great potential in capturing the relationship between objects, tasks, and grasp types as they outperform all other models. Achieving a micro f1 score of 0.774 while distinguishing between over 40 tasks, there is a difference of less than 0.02 to existing works that do not consider this information. As a result, these findings may contribute to developing methods that further mitigate arbitrariness during grasp choice. We are interested in further investigating how to model object manipulation using more sophisticated text representations and hope to see a stronger emphasis on this aspect in the grasp classification literature.

Acknowledgements. This work is supported by the German Federal Ministry of Education and Research as a part of CAMELOT (grant no. 01IW20008) and MultiGrasp (grant no. 01IS23064).

References

1. Bullock, I.M., Feix, T., Dollar, A.M.: Finding small, versatile sets of human grasps to span common objects. In: 2013 IEEE International Conference on Robotics and Automation. pp. 1068–1075. IEEE (2013). https://doi.org/10.1109/ICRA.2013.6630705
2. Bullock, I.M., Zheng, J.Z., De La Rosa, S., Guertler, C., Dollar, A.M.: Grasp frequency and usage in daily household and machine shop tasks. IEEE Transactions on haptics **6**(3), 296–308 (2013). https://doi.org/10.1109/TOH.2013.6
3. Cutkosky, M.R.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Transactions on robotics and automation **5**(3), 269–279 (1989)
4. Das, A., Chattopadhyay, A., Alia, F., Kumari, J.: Grasp-Pose Prediction for Hand-Held Objects. In: Emerging Technology in Modelling and Graphics, pp. 191–202. Springer (2020). https://doi.org/10.1007/978-981-13-7403-6_19
5. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)
6. Deng, Z., Fang, B., He, B., Zhang, J.: An adaptive planning framework for dexterous robotic grasping with grasp type detection. Robotics and Autonomous Systems **140**, 103727 (2021). https://doi.org/10.1016/j.robot.2021.103727
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423
8. Feix, T., Bullock, I.M., Dollar, A.M.: Analysis of human grasping behavior: Correlating tasks, objects and grasps. IEEE Transactions on haptics **7**(4), 430–441 (2014). https://doi.org/10.1109/TOH.2014.2326867
9. Feix, T., Bullock, I.M., Dollar, A.M.: Analysis of human grasping behavior: Object characteristics and grasp type. IEEE Transactions on haptics **7**(3), 311–323 (2014). https://doi.org/10.1109/TOH.2014.2326871
10. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. IEEE Transactions on human-machine systems **46**(1), 66–77 (2015). https://doi.org/10.1109/THMS.2015.2470657

11. Ghazaei, G., Alameer, A., Degenaar, P., Morgan, G., Nazarpour, K.: An exploratory study on the use of convolutional neural networks for object grasp classification. In: 2nd IET International Conference on Intelligent Signal Processing 2015 (ISP). pp. 1–5 (2015). https://doi.org/10.1049/cp.2015.1760
12. Hung, L.P., Alias, S.: Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. Journal of Advanced Computational Intelligence and Intelligent Informatics **27**(1), 84–95 (2023). https://doi.org/10.20965/jaciii.2023.p0084
13. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. pp. 427–431. Association for Computational Linguistics (2017)
14. Kleer, N., Feick, M., Feld, M.: Leveraging publicly available textual object descriptions for anthropomorphic robotic grasp predictions. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7476–7483. IEEE (2022). https://doi.org/10.1109/IROS47612.2022.9981541
15. Koubaa, A.: Rosgpt: Next-generation human-robot interaction with chatgpt and ros. Preprints.org p. 2023040827 (2023). https://doi.org/10.20944/preprints202304.0827.v2
16. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International conference on machine learning. pp. 1188–1196. PMLR (2014)
17. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
18. Lu, Q., Hermans, T.: Modeling grasp type improves learning-based grasp planning. IEEE Robotics and Automation Letters **4**(2), 784–791 (2019). https://doi.org/10.1109/LRA.2019.2893410
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
20. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
21. Neelakantan, A., Shankar, J., Passos, A., McCallum, A.: Efficient non-parametric estimation of multiple embeddings per word in vector space. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1059–1069. Association for Computational Linguistics (2014). https://doi.org/10.3115/v1/D14-1113
22. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014). https://doi.org/10.3115/v1/D14-1162
23. Rao, A.B., Krishnan, K., He, H.: Learning robotic grasping strategy based on natural-language object descriptions. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 882–887. IEEE (2018). https://doi.org/10.1109/IROS.2018.8593886
24. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
25. Selva Birunda, S., Kanniga Devi, R.: A review on word embedding techniques for text classification. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020 pp. 267–281 (2021)