Leveraging Publicly Available Textual Object Descriptions for Anthropomorphic Robotic Grasp Predictions

Niko Kleer, Martin Feick, and Michael Feld*

***Abstract*— Robotic systems using anthropomorphic end-effectors face tremendous challenges choosing a suitable pose for grasping an object. The fact that the choice of a grasp is influenced by the physical properties of an object, the intended task, and the environment results in a considerable amount of variables. The majority of models targeted towards enabling such robots to determine a suitable grasping pose rely on computer vision techniques, sometimes complemented by textual data. This paper investigates the potential of publicly available textual descriptions to predict a suitable grasping pose for anthropomorphic end-effectors. To this end, we have retrieved textual descriptions from Wikipedia, Wiktionary, and WordNet as well as a number of well-known dictionaries for 100 everyday objects. We analyze and compare the prediction quality of multiple learning methods while showing that a support vector machine-based approach can utilize this data for achieving a prediction accuracy above 0.75. Finally, we make our collected data available to the research community.**

I. INTRODUCTION

The execution of many tasks requires robotic systems to learn about the representation of the objects they interact with and how to handle them. In order to establish such representations, it is usually necessary to incorporate models that are based on a considerable amount of data. One of the most common tasks in robotics that requires knowledge about the physical properties of objects is robotic grasping. In particular, robots using anthropomorphic end-effectors that are expected to grasp objects in a similarly dexterous manner as humans face numerous challenges (see Figure 1). This is because, as opposed to a parallel jaw gripper, an anthropomorphic hand enables more fine grained control, but also requires additional knowledge about the object. More specifically, literature in the field of human grasp analysis emphasizes the influence of an object's physical properties, the intended task, and the environment on the choice of a suitable grasping pose [1], [2], [3]. Considering all these factors yields a high complexity with regards to the choice of a suitable grasping pose for a robotic system. Therefore, we need to develop methods that are capable of making sensible predictions based on the aforementioned criteria.

Computational models for predicting a suitable pose for grasping an object often exclusively rely on computer vision techniques [4], [5], [6], [7], [8]. Similarly, such approaches can be utilized for the automatic choice of a grasp in the field of prosthesis [9], [10]. Besides visual features, textual data also represents a rich source of information for robots to learn about the physical properties of objects.

*All authors are with the DFKI, Saarland Informatics Campus, Saarland University, 66123 Saarbrücken, Germany {niko.kleer;martin.feick;michael.feld}@dfki.de

Fig. 1. Demonstrates conceptually how a Pepper robot successfully learns how to grasp an unknown object using an anthropomorphic end-effector.

Therefore, textual descriptions extracted from sources such as Wikipedia or WordNet may complement computer vision approaches in order to establish a symbolic representation of an object's physical properties [11], [12]. Furthermore, the inclusion of textual data may also enable generating explanations in case of problematic situations, allow the use of semantic ontologies, or bridge the gap to natural human-robot communication.

In order to understand the contribution of textual data to the challenge of grasp pose predictions, we explore only this single modality in the context of publicly available textual object descriptions. To our knowledge, the exclusive use of textual object descriptions has not been published on by other work except for [13] for this purpose. Our work makes the following contributions:

- Based on publicly available textual object descriptions, we analyze and compare the prediction quality of multiple learning models and show that a support vector machine-based approach consistently yields a prediction accuracy above 0.75.
- For the systematic annotation of our retrieved descriptions, we have conducted a crowdsourcing study in which participants were asked to choose the most suitable pose for grasping 100 everyday objects.
- To support research in this field, we provide a dataset containing our retrieved object descriptions and the results of our study[1].

The paper is structured as follows. In the next section, we

[1]https://github.com/nikleer/TextBasedGrasps

provide an overview of the relevant literature regarding this work. After that, we elaborate on our procedure for retrieving publicly available textual object descriptions. In Section IV, we describe each aspect of our grasp pose learning approach explaining our selection of grasp poses, data annotation procedure and providing a detailed learning model analysis. Moving on, we discuss our investigations in Section V. Finally, the last section presents our conclusion.

II. RELATED WORK

The literature related to this work can be divided into two fields. First, it is important to understand the grasp types from the field of human grasp analysis as we have to sensibly choose a set of grasps that represent the target classes of our prediction model. Second, related work from the field of grasp prediction models provides insights into earlier approaches that aimed to overcome a similar or the same challenge as this work.

A. Analysis of Human Grasps

It was Napier [14] who first distinguished between the so-called precision and power-grip. At the same time, he discussed the various factors that influence the choice of a grasp. Incorporating Napier's terminology, Cutkosky [1] provided a hierarchical taxonomy of human grasps that set the foundation for more systematic research towards what could be considered a complete taxonomy of human grasps. In fact, Feix et al. [2] later presented the so-called GRASP taxonomy, the most complete and comprehensive taxonomy of human grasps currently known. Their work is based on the observation of grasp frequencies as well as the influence of an object's properties and the corresponding task on the choice of a grasp [15], [16], [17]. In contrast to their quantitative approach, Stival et al. [3] investigated the systematic structuring of human grasps from a qualitative point of view by retrieving electromyographic and kinematic data. By determining the similarity of grasping poses based on their data, they present a hierarchical structuring of grasps for each modality and merge all their results into one joint hierarchy.

For certain applications in the field of robotics and prosthesis, the literature outlined in this section enables simplifying predicting a suitable grasping pose to a large extent. Therefore, such taxonomies serve as a basis for grasp prediction models that are discussed subsequently.

B. Grasp Prediction Models

A considerable number of models targeted towards determining a suitable pose for grasping an object are based on computer vision techniques. Numerous publications have investigated the accuracy that convolutional neural networks can achieve when trained on a large collection of object-based images [4], [6], [7], [8]. However, especially Yang et al. [5] emphasize that such approaches should also focus on taking the intended action of a grasp into account, which is often neglected. This factor, as well as task-dependent constraints, are crucial for the automatic determination of a grasp in prosthesis applications [9], [10]. For more complex operations, such as task-oriented grasping, robots incorporate more sophisticated strategies for grasping an object [18], [19]. Still, such systems usually make use of computer vision approaches. To the best of our knowledge, the only case of a grasp prediction model exclusively based on textual descriptions was presented by Rao et al. [13], [20]. It is important to note that the authors based their prediction model on manually generated descriptions. Furthermore, their descriptions follow a specific structure parsed through regular expressions and contain precise information about the size of each object. In this work, however, we are interested in investigating the potential of publicly available textual object descriptions for predicting a suitable grasping pose.

In the following section, we introduce our methodology for establishing a corpus of publicly available object descriptions that serves as a basis for the learning models investigated in Section IV-D.

III. RETRIEVING OBJECT DESCRIPTIONS

In order to investigate the extent to which publicly available textual object descriptions can be used for predicting a suitable grasping pose, we first require an appropriate corpus. For such a corpus, it would be desirable to obtain descriptions that focus on describing the general shape and purpose of an object since these factors strongly contribute towards the choice of a grasp. However, in terms of freely accessible data sources, our options are limited. Below, we provide a description for each source of textual data that we have used for establishing our corpus.

- **Wikipedia**: Clearly, Wikipedia represents a highly popular source of information all over the globe as the current size of the English Wikipedia encompasses nearly 6.5 million articles. We gather textual object descriptions from Wikipedia by, first, pre-processing an XML-dump[2] as the raw data contains a considerable amount of annotations that make the automatic filtering of clean textual descriptions cumbersome. To this end, we use the WikiExtractor [21] Python script which allows us to extract clean textual passages from all articles in the XML-dump and turn them into the JSON format. As a result, the data can easily be interpreted and processed by an algorithm. It is important to note that Wikipedia articles contain a large amount of more general information (e.g. history-related facts) that does not provide an explicit value for predicting a suitable pose for grasping an object. Therefore, we only extract the abstract as it is most related to the main characteristics of an object.
- **Wiktionary**: In contrast to Wikipedia, which commonly describes objects in a lengthy manner, the Wiktionary provides shorter and more concise descriptions with regards to the characteristics of an object. For example, Wikipedia describes the object "bottle" by using a total of 49 words whereas the Wiktionary only uses 18.

[2] https://dumps.wikimedia.org/backup-index.html

TABLE I

GENERAL STATISTICS FOR OUR RETRIEVED TEXTUAL OBJECT DESCRIPTIONS FOR EACH SOURCE.

	Wikipedia	Wiktionary	WordNet	Collins	Merriam	Macmillan	American	Lexico	Dictionary
Descriptions	89	98	77	92	85	88	87	91	90
Average Length	72.4	15.3	12.4	21.3	15.1	17.6	22.0	18.7	20.9
Number of Words	6449	1509	956	1961	1284	1556	1916	1708	1884

Extracting these object descriptions also turns out to be easier as there already exists a machine-readable version of the Wiktionary[3] in the JSON format.

- **WordNet**: WordNet represents a lexical database structuring English terms into so-called synsets [22]. Synsets that share semantic or lexical relations are interlinked with each other. Furthermore, WordNet provides a generally short and concise definition for each synset. For our purpose, we specifically extract these definitions for all objects of our interest using the well-known Natural Language Toolkit (NLTK) Python library [23].
- **Other Dictionaries**: In order to further augment our corpus with more data about an object's physical properties and its usage, we additionally retrieve textual descriptions from a number of established dictionaries. More specifically, we included Collins' Dictionary[4], Merriam-Webster's Dictionary[5], the Macmillan Dictionary[6], the American Heritage Dictionary[7] as well as definitions from Lexico[8] and dictionary.com into our retrieval process. It is worth mentioning that the data of each dictionary can be accessed via a dedicated API.

We have used the above sources in order to extract textual descriptions for a total of 100 everyday objects including, for example, fruit (e.g. apple and pear), tools (e.g. scissors and screwdriver), and a number of different writing and eating utensils (e.g. ballpoint pen and fork). We follow a simple methodology by extracting the definition of the first noun that matches our desired object. After that, the extracted text of each definition is pre-processed (see Section IV-D). Table I provides an overview of general statistics with regards to our retrieved data for each source, including the total number of descriptions, their average length, and the total number of words. For space-related reasons, longer dictionary names are shortened by referring to them using only one word. The reason why we were not able to retrieve a description for all objects from all sources is that there exist cases where an object cannot be found, such as the object "chess piece". In other cases, different terms are used for referring to the same object (e.g. "sheet of paper" versus "paper").

The object descriptions we have retrieved serve as a basis for our learning model analysis. In the next section, we elaborate on our grasp pose learning approach.

[3]https://kaikki.org/dictionary/index.html
[4]https://www.collinsdictionary.com
[5]https://www.merriam-webster.com
[6]https://www.macmillandictionary.com
[7]https://www.ahdictionary.com
[8]https://www.lexico.com

IV. GRASP POSE LEARNING APPROACH

Before we can analyze and compare the prediction quality of multiple established learning methods, there are two aspects specifically that we need to discuss. We need to (a) agree on a set of grasp poses to classify, and (b), since there is no ground truth available for this challenge, develop a systematic annotation procedure for them. We subsequently start by addressing these aspects. After that, we move on to our detailed learning model analysis.

A. Selection of Grasp Poses

Sensibly selecting grasp poses for annotating objects with regards to their most suitable grasp represents an important factor. Choosing too many grasp classes increases the likelihood of introducing confusion, i.e. the most appropriate grasp becomes less clear. On the other hand, considering too few might eliminate the possibility of grasping certain objects completely. In order to determine the smallest number of grasps that spans over as many objects as possible during various tasks, we based our decision on the statistical observations from the literature [16], [17], [24]. Considering these observations, the grasping poses most distinguishable that span over the largest set of objects are referred to as medium wrap, lateral, tripod, and writing tripod. These grasping poses also result in a perfect split between power- (lateral and medium wrap) and precision grasps (tripod and writing tripod) [1]. It is worth mentioning that the same methodology was used by Salvado [7] who chose nearly the same grasping poses for the annotation of images. Our choices also appear sensible in accordance with the quantitative grasp taxonomy established by Stival et al. [3] as we choose exactly one grasp from each category, excluding ring grasps.

B. Data Annotation Procedure

Due to the factors influencing the choice of a suitable grasp, systematically labeling a dataset represents a challenging task. We have found that, in the literature, many authors do not provide details regarding their data annotation approach or follow an elimination process [4], [6], [7], [9]. For overcoming this problem, we have conducted a crowdsourcing study on the crowdsourcing marketplace Amazon Mechanical Turk (MTurk)[9]. The platform enables the automatic annotation of large amounts of data by distributing the annotation task to users of the platform that can be located all over the globe while jointly working on the same task. At the same time, study participants (so-called Workers) gain a small amount of money for each annotated sample.

[9]https://www.mturk.com

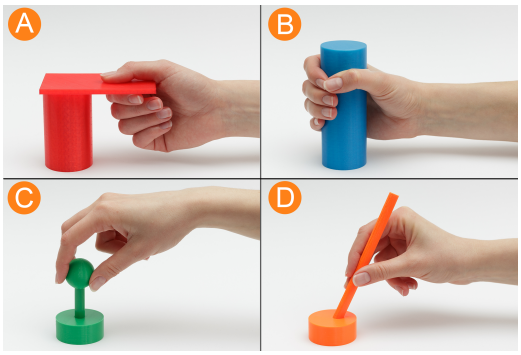

Fig. 2. The grasping poses referred to as lateral (A), medium wrap (B), tripod (C), and writing tripod (D). Figure adapted from [25]

TABLE II

TEN MOST FREQUENT TERMS FOR EACH GRASP POSE CLASS IN OUR DATA. GRASPING POSES ARE ABBREVIATED FOR SPACE REASONS.

Lateral		Medium		Tripod		Writing	
w	*c(w)*	*w*	*c(w)*	*w*	*c(w)*	*w*	*c(w)*
card	48	fruit	73	small	30	pen	32
paper	37	tree	38	piece	29	blade	32
hair	28	handle	36	tree	22	handle	31
container	25	container	33	nut	20	small	27
phone	17	small	32	fruit	18	cutting	26
plastic	16	skin	31	genus	17	consisting	26
bank	15	yellow	30	metal	13	instrument	22
money	15	cup	28	ear	13	brush	21
small	15	plant	27	place	12	tube	20
food	15	large	27	plastic	12	metal	19

In the crowdsourcing study we have conducted, study participants were asked to choose the most suitable grasping pose for **holding an object**. To make this decision, participants were shown an image of each grasping pose we have mentioned in the previous section (see Figure 2). Our visualisations demonstrate the general kinematics of each grasp at the example of a 3D-printed prop that is kept as neutral as possible in order to not resemble a real object [25]. Participants were given a maximum of five minutes to choose the most suitable grasping pose. It was not mandatory for a participant to make a choice. Overall, we gathered 20 assessments for each object, resulting in a total of 2000 assessments.

C. Study Results

Our study results show that participants have determined the lateral to be the most suitable grasp for 27 objects, the medium wrap for 28 objects, the tripod for 26 objects, and the writing tripod in case of 24 objects. Participants were unable to distinctly determine one most suitable grasping pose for five objects, resulting in a tie. Another aspect to note is that, for 74 objects, the most suitable grasping pose was determined by an absolute majority (i.e. the grasping pose received more than ten votes). This is an aspect that we have considered during our learning model analysis as we will discuss in the next section. Furthermore, by annotating our data according to these results, we are able to gain more insights regarding the most frequent words prevalent in each class as shown in Table II where we denote a word by *w* and its corresponding frequency by *c(w)*. Note that these statistics fully ignore stopwords and, for the sake of simplicity, we did not include the five objects that could not clearly be assigned to only one grasping pose. By examining the word frequencies, we can observe a reasonably strong association between a grasping pose and objects typically grasped using the respective pose (e.g. a card using the lateral or a pen using the writing tripod). Finally, there is one more notable aspect to point out. As we are able to systematically annotate our data, each label enables us to infer whether an object is supposed to be grasped using a power- or precision grasp, as briefly discussed in Section IV-A. We may leverage this fact to further augment our textual data by adding a one-word description (i.e. either power or precision) to each sample.

Based on the results of our study, we move on to our detailed learning model analysis in the next section.

D. Analysis of Grasp Prediction Models

This section provides an analysis and comparison of multiple classification models for making text-based grasp pose predictions. To this end, we have chosen and evaluated a total of four classification approaches, namely Naïve Bayes, the probabilistic document ranking algorithm BM25+, a word embedding-based convolutional neural network, and a support vector machine. Through these choices, we cover a large space of established, powerful, and diverse methods. We consider it most sensible to have a closer look at the predictions of each individual model on an object-level basis as this allows us to gain a better understanding with regards to where each approach fails to classify objects correctly. This allows us to investigate if particular approaches appear more suitable for classifying specific types of objects.

All textual object descriptions in our dataset were pre-processed by applying stopword filtering, lemmatization, and stemming. Furthermore, all descriptions for an object were concatenated. For obtaining the prediction accuracy of our models, we applied a stratified 10-fold-cross-validation to our pre-processed data. For all objects that were assigned a grasp pose label based on an absolute majority (i.e. the grasp pose was considered most suitable in **more than half** of all cases by study participants), a prediction must be equivalent to the assigned label to be considered correct. In all the other cases, a prediction must be equivalent to the grasp pose determined **most or second most suitable by study participants**. We enforce this evaluation methodology in order to deal with the uncertainty in the determination of a suitable grasping pose for the 24 objects that have not been assigned a label based on an absolute majority.

E. Analysis Results

We elaborate on the performance of each model while supporting our findings with tables that list examples for significant differences in the classification of specific objects between models.

1) Naïve Bayes: Starting with Naïve Bayes, this machine learning model represents the classifier with the worst performance. As shown in Table III, the model fails to classify numerous objects that are otherwise classified correctly by the majority of the other models. For the

TABLE III
SIGNIFICANT DIFFERENCES IN THE CLASSIFICATION OF OBJECTS WITH REGARDS TO OUR NAÏVE BAYES CLASSIFIER. CONSISTENTLY CLASSIFYING AN OBJECT CORRECTLY IS MARKED BY ✓, UNCERTAINTY INDICATED THROUGH ○ AND CONSISTENT FALSE CLASSIFICATION BY ×.

Object	Naïve Bayes	SVM	BM25+	CNN
Bead	○	✓	✓	○
Fork	×	✓	✓	✓
Ladle	×	✓	✓	✓
Mobile Phone	○	✓	✓	✓
Paintbrush	×	✓	×	✓
Potato Chip	×	✓	×	✓
Vase	○	✓	✓	✓

objects ”fork” and ”paintbrush”, the model predicts the lateral instead of the writing tripod. In case of the object ”ladle”, which received an even distribution of votes between the lateral and writing tripod, our model predicts the medium wrap. Clearly, our Naïve Bayes classifier often fails to capture the distinct textual features that contribute towards correctly classifying an object. Nevertheless, the model is able to consistently classify a total of 69 objects correctly and **its prediction accuracy converges to 0.69**.

2) BM25+: The BM25+ represents a variation of the BM25 document ranking algorithm and assigns a relevance score to documents based on a given query [26]. In our case, the training data is used for establishing a grasp specific corpus (i.e. a document) for each class. For each sample in the test data (i.e. the queries), the BM25+ assigns a relevance score to each grasp-specific corpus which allows us to make a statement about their individual relevance. Similar to our Naïve Bayes classifier, the BM25+ sometimes fails to classify objects that are otherwise classified correctly by the majority of the other models as shown in Table IV. Once again, the model fails to classify the objects

TABLE IV
SIGNIFICANT DIFFERENCES IN THE CLASSIFICATION OF OBJECTS WITH REGARDS TO OUR BM25+ CLASSIFIER.

Object	Naïve Bayes	SVM	BM25+	CNN
Comb	✓	✓	×	○
Paintbrush	×	✓	×	✓
Potato Chip	×	✓	×	✓
Toothbrush	✓	✓	×	✓
Chestnut	✓	✓	×	✓

”paintbrush” and ”potato chip” correctly, assigning the highest relevance score to the lateral and medium wrap respectively. Moreover, the BM25+ is the only model that fails to classify the objects ”toothbrush” and ”chestnut” correctly. We further did not observe a single case where this approach yields better results for an individual object when compared to the others. In total, the model is able to consistently classify a total of 69 objects correctly and **its prediction accuracy converges to 0.7**.

3) Convolutional Neural Network: For our convolutional neural network, we have determined a simple architecture comprising a word embedding-based input layer followed by two layers of convolution, one layer of global max-pooling, and three dense layers to converge to the highest prediction accuracy. For our embedding layer, we use 100-dimensional pre-trained[10] GloVe embeddings [27]. Apart of our final dense layer, which uses sigmoid activation, all layers utilize the Rectified Linear Unit (ReLU) activation function. Finally, our model is trained using the adam optimizer.

Based on the above described architecture, in contrast to Naïve Bayes and the BM25+, we found that this approach sometimes manages to classify objects correctly where at most one of the others is capable of successfully predicting a suitable grasp. At the same time, a few objects that are usually assigned their correct grasp label by the other models cannot consistently be classified. Table V provides examples for both cases. It appears that the influence

TABLE V
SIGNIFICANT DIFFERENCES IN THE CLASSIFICATION OF OBJECTS WITH REGARDS TO OUR CNN CLASSIFIER.

Object	Naïve Bayes	SVM	BM25+	CNN
Dice	×	○	×	✓
Key	×	×	○	✓
Spatula	×	×	×	✓
Cup	✓	✓	✓	○
Earring	✓	✓	✓	○
Scissors	✓	✓	✓	○
Screwdriver	✓	✓	✓	○

of our embedding layer causes a considerable amount of inconsistencies with regards to the classification of individual objects as compared to the other approaches. Objects such as ”dice”, ”key”, and ”spatula” that are usually correctly classified by at most one of our approaches do not represent an issue for the convolutional neural network. At the same time, the embedding also seems to contribute to introducing a significant amount of confusion for objects like ”cup”, ”earring”, ”scissors”, and ”screwdriver”, which can be classified correctly by all the other models. Even though this is the case, the model is able to consistently classify a total of 72 objects correctly while **converging to a prediction accuracy of 0.75**.

4) Support Vector Machine: Finally, the support vector machine represents our best classifier. Our model is based on a tf-idf (term frequency-inverse document frequency) matrix where each object’s tf-idf vector is used for training. It is important to note that the cut-off hyperparameter is set to two (i.e. all words appearing less than twice are ignored)

[10]https://nlp.stanford.edu/projects/glove

and the support vector machine uses a linear kernel. We have observed that this model is capable of correctly classifying nearly all objects that have been assigned the correct class label by at least one other classifier. Moreover, there are numerous cases of uncertainty as opposed to falsely classifying objects completely as indicated by Table VI. In

TABLE VI

SIGNIFICANT DIFFERENCES IN THE CLASSIFICATION OF OBJECTS WITH REGARDS TO OUR SVM CLASSIFIER.

Object	Naïve Bayes	SVM	BM25+	CNN
Bowl	×	○	×	×
Bun	×	○	×	○
Chocolate Bar	×	○	×	×
Sponge	○	○	×	×
Spoon	×	✓	×	○

general, our support vector machine classifier outperforms all the other models, sometimes by a significant margin. The approach enables us to consistently classify a total of 77 objects correctly and **converges to a prediction accuracy of 0.79**.

In addition to the selected examples that we have used for demonstrating strengths and weaknesses in the classification of specific objects, a complete overview can be found in the Table VII and Table VIII. Overall, a total of 13 objects was consistently misclassified by all models. For gaining insights with regards to which grasp poses could be predicted more easily by our models, we have also obtained their respective confusion matrices. Figure 3 shows the confusion matrices of our classifiers. It is important to note that we have excluded all objects where a model's prediction fluctuated too much between multiple classes. More specifically, this resulted in us excluding five, two, five, and twelve objects in case of our Naïve Bayes, BM25+, SVM, and CNN classifier respectively.

Following up on our investigations, we move on to discussing several notable aspects in the next section.

V. DISCUSSION

In order to leverage the online extraction of publicly available textual object descriptions in robotic grasping applications, it would be desirable to disambiguate terms with multiple meanings (homonyms). This is because, in a few cases, we retrieve textual descriptions that do not refer to the desired object. One example for such an object is "sponge". Since a sponge is not only an everyday object for washing but an underwater animal as well, some dictionaries list the latter as the first noun for the term sponge. We assume that the incorporation of disambiguation methods could have been advantageous and should enable the retrieval of textual descriptions for an even higher number of objects.

Regarding our results, it is evident that especially objects labeled with the tripod grasp contribute towards lowering the prediction accuracy of our models. All confusion matrices visualized in Figure 3 clearly show the difficulty that lies in distinguishing those objects from the others. We assume that one reason for this is the seemingly arbitrary usage of

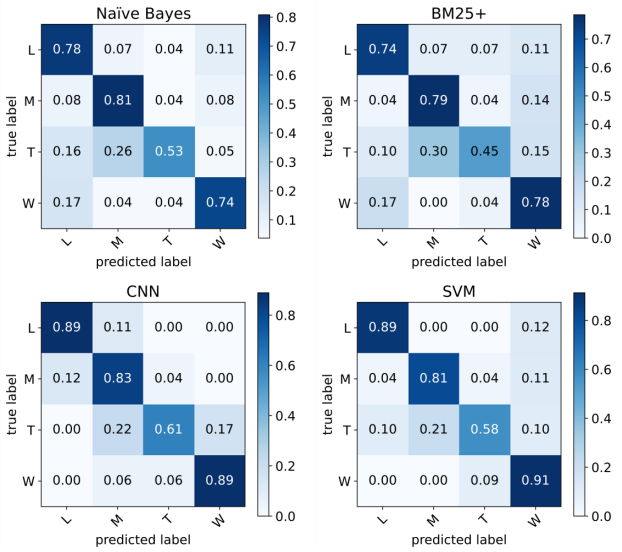

Fig. 3. Confusion matrix for each learning model. For space reasons, we refer to each grasping pose by its first letter.

words in our extracted data that describe an object's size. In particular, the word "small" represents the most frequent word in the tripod grasp pose class. This seems sensible from a logical perspective as the grasping pose is usually applied to smaller objects. At the same time, the word shows a high frequency in all the other grasp pose classes, therefore resulting in a confusing or even improper assignment of words to a number of objects. Another reason is likely the fact that our data contains numerous fruit that are distributed over the tripod and medium wrap grasp classes. As the distribution, however, is strongly biased towards the latter class, there may not be enough feature information for a clear separation. This assumption is supported by the fact that the highest number of false predictions arises in exactly this case. A concrete example of a fruit where this effect can be observed is the object "strawberry", which cannot be classified correctly by any of our models.

It should be noted that the prediction accuracies reported in the previous section could be considered a lower bound for each model. This is because our data contains a few objects where the second most suitable grasp still represents a valid choice even though the most suitable grasp was determined by an absolute majority in our study. For example, eleven participants determined the writing tripod to be the most suitable grasp for holding a spoon. At the same time, six participants chose the lateral grasp instead. Both of these grasping poses could be applied in different task-dependent contexts (e.g. using the writing tripod for eating and the lateral for simply holding the object or during a handover). We did not consider such special cases during our evaluation as it would have eliminated the systematic nature of our methodology. Establishing a grasp prediction model that considers a higher number of tasks and the inclusion of contextual information might help to resolve the problem of ambiguous grasp pose selection.

VI. CONCLUSIONS

In this paper, we have thoroughly investigated the potential of publicly available textual object descriptions for the purpose of anthropomorphic grasp predictions. To this end, we have elaborated on our procedure for obtaining the descriptions, our methodology for systematically labelling our data, and our analysis of multiple learning models. We have shown that such approaches yield great potential as our support vector machine-based classifier achieves a prediction accuracy above 0.75. Moreover, we make our retrieved textual object descriptions and the results of our conducted study available to the research community.

Our results indicate that grasp pose predictions based on unstructured textual resources may represent an alternative or an extension to other modalities. Moreover, publicly available textual resources seem to serve for applications such as ad-hoc retrieval of descriptions for new (i.e. unknown) objects. We would like to further investigate the possibilities to improve such models while applying them in the context of human-robot interactive situations where, for example, a human teaches a robot how to grasp an object by providing a verbal description. Finally, the inclusion of quantitative data such as as an object's size, which is partially examined in the literature, also appears promising and we aim to explore how such data can be incorporated using intuitive methods.

ACKNOWLEDGMENT

This work is supported by the German Federal Ministry of Education and Research (grant no. 01IW20008) as a part of CAMELOT - Continuous Adaptive Machine-Learning of Transfer of Control Situations.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, and K. Nazarpour, "An exploratory study on the use of convolutional neural networks for object grasp classification," *2nd IET International Conference on Intelligent Signal Processing 2015 (ISP)*, pp. 1–5, 2015.

[5] 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.

[6] G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, and K. Nazarpour, "Deep learning-based artificial vision for grasp classification in myoelectric hands," *Journal of neural engineering*, vol. 14, no. 3, p. 036025, 2017.

[7] F. Lluís Salvadó, "Grasp prediction with convolutional neural networks," B.S. thesis, Universitat Politècnica de Catalunya, 2017.

[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] 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.

[10] 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.

[11] T. Nguyen, N. Gopalan, R. Patel, M. Corsaro, E. Pavlick, and S. Tellex, "Robot object retrieval with contextual natural language queries," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, 2020.

[12] D. Nyga, M. Picklum, and M. Beetz, "What no robot has seen before — probabilistic interpretation of natural-language object descriptions," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 4278–4285.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] T. Feix, I. M. Bullock, and A. M. Dollar, "Analysis of human grasping behavior: Object characteristics and grasp type," *IEEE transactions on haptics*, vol. 7, no. 3, pp. 311–323, 2014.

[18] 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.

[19] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Preparatory object reorientation for task-oriented grasping," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 893–899.

[20] A. B. Rao, H. Li, and H. He, "Object recall from natural-language descriptions for autonomous robotic grasping," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2019, pp. 1368–1373.

[21] G. Attardi, "Wikiextractor," https://github.com/attardi/wikiextractor, 2015.

[22] Princeton University, "About wordnet," 2010. [Online]. Available: https://wordnet.princeton.edu/

[23] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[24] I. M. Bullock, T. Feix, and A. M. Dollar, "Finding small, versatile sets of human grasps to span common objects," in *2013 IEEE International Conference on Robotics and Automation*. IEEE, 2013, pp. 1068–1075.

[25] M. Feick, K. P. Regitz, A. Tang, and A. Krüger, "Designing visuo-haptic illusions with proxies in virtual reality: Exploration of grasp, movement trajectory and object mass," in *CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–15.

[26] A. Trotman, A. Puurula, and B. Burgess, "Improvements to bm25 and language models examined," in *Proceedings of the 2014 Australasian Document Computing Symposium*, 2014, pp. 58–65.

[27] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

TABLE VII

THE FIRST HALF OF THE OBJECTS WE HAVE USED FOR LEARNING SUITABLE GRASPING POSES.

Object	Naïve Bayes	SVM	BM25+	CNN
Acorn	✓	✓	✓	✓
Apple	✓	✓	✓	✓
Apricot	✓	✓	✓	✓
Avocado	✓	✓	✓	✓
Ballpoint Pen	✓	✓	✓	✓
Banana	✓	✓	✓	✓
Banknote	✓	✓	✓	✓
Barrette	×	×	×	×
Battery	×	×	×	×
Bead	○	✓	✓	○
Beer Bottle	✓	✓	✓	✓
Billiard Ball	✓	✓	✓	✓
Book	✓	✓	✓	✓
Bookmark	✓	✓	✓	✓
Bottle	✓	✓	✓	✓
Bowl	×	○	×	×
Box	✓	✓	✓	✓
Bun	×	×	×	×
Candle	×	○	×	○
Chalice	✓	✓	✓	✓
Chalk	×	×	×	×
Chess Piece	✓	✓	✓	✓
Chestnut	✓	✓	×	✓
Chocolate Bar	×	○	×	×
Cigarette	✓	✓	✓	✓
Clementine	✓	✓	✓	✓
Coin	✓	✓	✓	✓
Comb	✓	✓	×	○
Compact Disk	✓	✓	✓	✓
Credit Card	✓	✓	✓	✓
Cucumber	✓	✓	✓	✓
Cup	✓	✓	✓	○
Dice	×	○	×	✓
Donut	✓	✓	×	✓
Earring	✓	✓	✓	○
Egg	✓	✓	✓	✓
Envelope	✓	✓	✓	✓
Flashlight	○	×	×	○
Flask	✓	✓	✓	○
Fork	×	✓	✓	✓
Fountain Pen	✓	✓	✓	✓
Frisbee	✓	✓	✓	✓
Gel Pen	✓	✓	✓	✓
Glasses	✓	✓	✓	✓
Glue Stick	×	×	×	×
Golf Ball	✓	✓	✓	✓
Hairbrush	×	×	×	×
Hammer	×	×	×	×
Handfan	✓	✓	✓	✓
Jar	✓	✓	✓	✓

TABLE VIII

THE SECOND HALF OF THE OBJECTS WE HAVE USED FOR LEARNING SUITABLE GRASPING POSES.

Object	Naïve Bayes	SVM	BM25+	CNN
Jug	✓	✓	✓	✓
Key	×	×	○	✓
Kiwifruit	×	×	×	×
Knife	✓	✓	✓	✓
Ladle	×	✓	✓	✓
Lemon	✓	✓	✓	✓
Lollipop	×	×	×	×
Marble	×	×	×	○
Marker Pen	✓	✓	✓	✓
Mobile Phone	○	✓	✓	✓
Mouse pad	✓	✓	✓	✓
Mug	✓	✓	✓	✓
Newspaper	✓	✓	✓	✓
Notepad	✓	✓	✓	✓
Onion	✓	✓	✓	✓
Paintbrush	×	✓	×	✓
Pan	✓	✓	✓	✓
Peach	✓	✓	✓	✓
Pear	✓	✓	✓	✓
Pencil	✓	✓	✓	✓
Pill	×	×	×	✓
Pliers	×	×	×	×
Plate	✓	✓	✓	✓
Playing Card	✓	✓	✓	✓
Potato	✓	✓	✓	✓
Potato Chip	×	✓	×	✓
Quill	✓	✓	✓	✓
Remote Control	✓	✓	✓	✓
Rubik's Cube	×	×	×	×
Scalpel	✓	✓	✓	✓
Scissors	✓	✓	✓	○
Screw	×	×	×	×
Screwdriver	✓	✓	✓	○
Sheet of Paper	✓	✓	✓	✓
Spatula	×	×	×	✓
Sponge	○	○	×	×
Spoon	×	✓	×	○
Strawberry	×	×	×	×
Syringe	✓	✓	✓	✓
Tennis Ball	✓	✓	○	✓
Thermometer	✓	✓	✓	○
Toothbrush	✓	✓	×	✓
Tube	✓	✓	✓	✓
Tweezers	✓	✓	✓	✓
Vase	○	✓	✓	✓
Wallet	✓	✓	✓	✓
Walnut	✓	✓	✓	✓
Whisk	✓	✓	✓	✓
Wine Bottle	✓	✓	✓	✓
Wine Glass	✓	✓	✓	✓