Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Object Recall from Natural-Language Descriptions for Autonomous Robotic Grasping

Achyutha Bharath Rao, Hui Li, and Hongsheng He*

Abstract— **Humans acquire grasping skills through repeated interaction with the objects; and they internalize the knowledge of various physical attributes of such objects. Even blindfolded, humans can reasonably estimate a suitable grasp pose given only the object's description. Human brain relies on the knowledge of having seen such objects and recalling its physical features such as shape, size, weight to estimate feasible grasps. In such a scenario, knowledge of an object's features is key to executing a successful grasp. This paper aims to provide this 'recall' ability to robots by proposing a methodology to represent human memory of objects with a dataset of objects and their physical features. A joint probability distance metric is derived, which can match the natural language descriptions to reference object features so as to recall and identify a particular object from the reference dataset, thereby facilitating better grasp planning. Experiments were performed to evaluate the accuracy of object recall, and simulation of an anthropomorphic robot hand was conducted for object grasping based on the recalled object features. The experiment results showed the accuracy of the proposed metric and effectiveness in object grasping.**

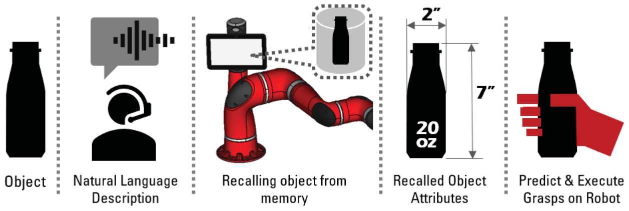

Object
Natural Language Description
Recalling object from memory
2″
7″
20 oz
Recalled Object Attributes
Predict & Execute Grasps on Robot

Fig. 1

IDENTIFYING THE OBJECT BASED ON ITS DESCRIPTION: FOR A SELECTED OBJECT, THE SYSTEM RECALLED THE OBJECT FROM ITS MEMORY AND EXTRACT THE FEATURES OF THAT OBJECT BASED ON A DESCRIPTION GIVEN BY USER. ACCORDING TO THE EXTRACTED FEATURES, GRASP STRATEGIES ARE GENERATED.

I. INTRODUCTION

Autonomous robotic hands with even a small fraction of human dexterity would greatly improve productivity and safety in several industrial and household application contexts. They can be used in unfamiliar hazardous environments to interact with objects and perform mechanical tasks. However, robotic grasping, especially with five-fingered robotic hands, is still evolving and as such the grasping abilities of robots are not comparable to human grasping skills today.

Humans have the most sophisticated grasping skills with a significant portion of the brain's sensorimotor apparatus dedicated to grasping. Behind this simple task of grasping an object, the human brain is executing a series of sub-tasks with the associated decision-making and error-correction process in real time. To complete each of these sub tasks the brain selects and executes appropriate motor strategy learned earlier by the human sensorimotor apparatus. These 'learned strategies', also called as action-phase controllers [6], utilize the input sensory signals and corresponding predictions by the nervous system to produce motor commands to accomplish the given motor task. To accurately estimate the specific motor output required, action-phase controllers need the information about an object's physical properties.

Learning techniques have been extensively used to solve object recognition, pose estimation, grasp planning and execution [7], [11]. Most of such research [8] was focused strongly on object recognition using images or 3D point clouds. Our research interest is focused on achieving similar objective but using only natural language descriptions. Object recognition is key to robot functioning and its efficacy in navigating the physical world. Especially for grasping, a much more refined information about the object is required. In addition to shape and size information, additional features such as an object's stiffness and fragility can make a difference in grasp strategy. Most common approach to gathering this information is by using 2D or 3D vision inputs [1], [5]. With our focus being on grasping using five fingered hand, we needed more than shape and size of the object. Furthermore, such features had to be extracted using natural language description of the object. Human expertise in recognizing objects from their description comes from our experiential learning. In emulating such a behavior by robots, we need to provide the robot an equivalent to human memory and experience and then use the natural language descriptions to find the closest match.

Human brain makes instantaneous decisions on grasp strategy, relying on past knowledge of having held or seen such

Achyutha Bharath Rao is with the Department of Industrial Systems and Manufacturing Engineering, Wichita State University, Wichita, KS, 67260, USA axbharathrao@shockers.wichita.edu.

Hui Li is with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, 67260, USA hxli4@shockers.wichita.edu.

Hongsheng He is with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS, 67260, USA hongsheng.he@wichita.edu.

*Correspondence should be addressed to Hongsheng He, hongsheng.he@wichita.edu.

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

objects and relying on our estimates of its physical features such as weight, texture, stiffness etc. Therefore, an effective system for robotic grasping should include a mechanism to perceive an object's physical attributes. Machine vision is widely used for robotic perception to recognize objects in a scene. But for grasping we need more than just the size and shape of objects. The objective of this paper is to use natural language descriptions to understand object affordances including shape, size, weight, texture, stiffness, etc., to enable a five-fingered anthropomorphic robotic hand to grasp the described target object.

In this paper, we propose a novel method to recognize objects using their natural language descriptions in order to equip robots with a similar prior knowledge of the objects, as shown in Fig. 1. Human descriptions of object features tend to be imprecise and incomplete. We therefore use the parsed object features and cross-reference them with a prior knowledge of the objects stored in the form of a dataset of various objects and their physical features. The object is identified based on calculated probabilities of the described features matching the object's features. Once the object is identified, the robot has complete access to the object's physical features and make better estimates of the suitable grasp type. In this approach, the object database models the human prior knowledge and experience, and the probabilistic search and identification represents the human memory recall and recognition process. Finding the closest match between the described and the recorded object features is treated as a problem of finding minimum vector distance between the features.

II. Description to Perception

The central objective of this study is to explore the idea that a robot can be equipped with skills to employ an appropriate grasping strategy to grasp an object, given only the natural language description of the target object. In general, the grasping task involves perceiving the object's physical features (affordances), predicting the appropriate grasp configuration, and executing the grasp by manipulating the robot's fingers. By early childhood, humans master these skills through repeated tactile interaction with the objects around. They gain and retain the knowledge about what to expect of the object in terms of shape, texture, weight, stiffness etc. Such understanding helps humans in crafting appropriate grasping strategies under various contexts. In this paper, we are attempting to explicitly impart these skills to the robot by employing learning models.

Asking a robot to choose a right grasp with just the natural-language description of an object is akin to asking a blindfolded person to grasp an object by describing the object features. It is easy to imagine that, even blindfolded, humans can easily accomplish this task especially if they are familiar with the object. Humans do this by drawing up on their experience of grasping same or similar objects – by first recalling the object's features and then choosing a grasp that is most suitable. Of course, this would be considerably more challenging if the subject has never seen or held the object being described.

Therefore, to estimate a suitable grasp, the robot has to be first equipped with the ontological knowledge of the object affordances and secondly the ability to identify the object based on its natural language descriptions. In the context of this study, the list of objects and their physical features would be the proxy for the ontological knowledge of the world, consisting of a limited number of graspable household objects. We propose to equip the robot with these skills by developing learning models to:

- ◇ Firstly, parse the natural-language description and extract as many of the physical features as possible of the object being described; and
- ◇ Secondly, we cross-reference this feature set with a dataset of objects, i.e., identifying the object with matching descriptions.

A. Prominent Object Attributes for Grasping

In choosing the features for our dataset, we planned to choose physical attributes of the object that significantly influence the grasping decisions. Grasp choices vary based on an object size, weight, shape and sometimes texture, stiffness fragility etc. [9]. Earlier studies [3], [4], [9] have found the following set of features shown in Table I are some of the most influential in making grasp choices in humans. They are the primary features that we look for in the descriptions of the objects while parsing for information.

TABLE I

Description of object feature sets.

Feature	Description	Value Range
a, b, c	Dimensions (cm) along orthogonal directions [3]	$a, b, c \in \mathbb{R}$ and $a \geq b \geq c$
m	Mass (grams)	mass $\in \mathbb{R}$
s	Shape classification per [2]	{thin, compact, prism, long, radial}
r	Rigidity of the object	{rigid, squeezable, floppy}
mt	Simplified material type description of the object.	{fabric, glass, metal, paper, plastic, rubber, wood, other}

B. Object Data Mining

Humans learn and familiarize the sizes, textures, weights of objects by interacting with the objects over many years spanning early childhood into adulthood. A robot need not be limited by the slow learning process. Provided that we give access to online storefronts, a robot can instantaneously scour such product websites and look for objects that match the natural language descriptions. In this paper, for standard evaluation of the proposed approach, we constructed an offline local database of common household objects and their physical attributes. In a way, this small dataset is representative of the collective online knowledge that one day robots can access to build their ontological worldview of objects and features, and the dataset can be conveniently expanded.

C. Translating Object Descriptions to Features

During our study, when asked to describe objects, it was natural for human subjects to describe the objects by stating its approximate dimensions such as “The object is about four inches in length and about three inches in diameter, weighs about ten ounces, and it is metallic.”. NLP techniques are used to parse such statements and extract object attributes that are needed to ‘recognize’ the object being described.

The natural language parsing applicable for this task is dealt in detail in the [12]. In summary, a combination of parts of speech tagging and chunking with regular expressions is used to parse specific feature descriptors as shown in 2. Using this approach, we extract as many of the features available in the description.

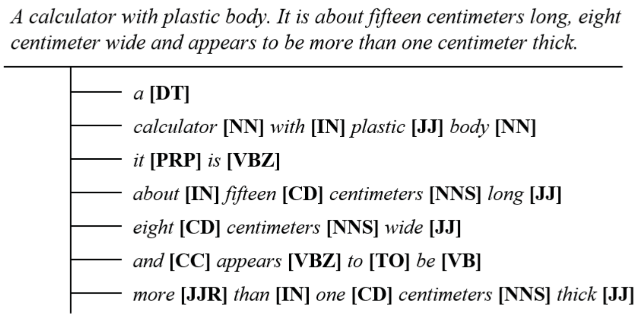

A calculator with plastic body. It is about fifteen centimeters long, eight centimeter wide and appears to be more than one centimeter thick.
a [DT]
calculator [NN] with [IN] plastic [JJ] body [NN]
it [PRP] is [VBZ]
about [IN] fifteen [CD] centimeters [NNS] long [JJ]
eight [CD] centimeters [NNS] wide [JJ]
and [CC] appears [VBZ] to [TO] be [VB]
more [JJR] than [IN] one [CD] centimeters [NNS] thick [JJ]

Fig. 2
EXAMPLE OF THE CHUNK TREE. THE INPUT STATEMENT AT THE TOP IS PARSED INTO CHUNKS AS SHOWN IN THE TREE USING REGULAR EXPRESSIONS.

III. OBJECT RECALL

Even with a reasonably detailed elucidation of an object’s features, human descriptions often tend to be either imprecise or incomplete or both. For example, the general tendency is to round-off dimensions and mass, often missing to mention certain features such as material type or texture or getting them wrong. A human subject can still work with the information available only because of the recall of having seen/held such an object.

In this paper, we propose to represent the human memory with a curated dataset of objects and their physical features. The feature set extracted from natural language description are represented as vectors with any categorical variables converted to one-hot encoded binary features, and a distance metric is proposed to compare the reference vector with each object in the dataset. The mapping of the target object to reference objects is measured by the distance metric of object descriptions.

In the natural language processing (NLP) domain, there are multiple popular distance metrics available to measure vectorial distances; however, these distance metrics did not yield the desired level of accuracy in the experiment. We therefore propose a distance metric derived based on joint probabilities to measure the vector distances. The rationale and the derivation for the distance metric are provided below.

Considering a dataset of objects and their physical features, let $f(i)$ represent the m features corresponding to an i^{th} object from this dataset

$$f_{(i)} = [f_1^{(i)}, f_2^{(i)}, f_3^{(i)} \dots f_j^{(i)} \dots f_m^{(i)}]$$

where $i{\in}[1, N]$ and $j{\in}[1, m]$. Let $f(o)$ represent the features parsed from the natural language description of the reference object and $f(o)$ is subset of $f(i)$

$$f(o) = [f_1^{(o)}, f_2^{(o)}, f_3^{(o)} \dots f_j^{(o)} \dots f_p^{(o)}] \text{ where } p \leq m.$$

Note that often, human descriptions do not contain all the feature descriptors. In such cases we need to make do with available information and compare only the available features.

In order to find a suitable match, we can use common distance measures such as Euclidean, Minkowski and Cosine distances measure which the proximity of a point in n-dimensional space. However, to increase the probability of the correct match, it was necessary to not only ensure the proximity of the points in the normed vector space, but also to ensure proximity of each individual feature. This would help in the identification accuracy and the confidence. To that end, we calculate the distance of each feature of the reference object with that of an object in the dataset. We then calculate the independent probability of the i^{th} object being selected conditioned on distance of the j^{th} feature. Features that are closer in value are assigned a higher probability.

Consider the feature $j \in [1, m]$ of an i^{th} object. We would like our metric to report a smaller value if the feature $f_j(i)$ is closer to $f_j(o)$. Let d_j be the absolute distance between the two feature values given by

$$d_j = |f_j(i) - f_j(o)|$$

Note that since not all features may be available in $f_j(o)$, only those features which exist both in $f_j(i)$ and $f_j(o)$ are used for calculating the distances.

We calculate the probability of the i^{th} object being mapped to the reference object given the distance d_j of the j^{th} feature as

$$P(i \rightarrow o|d_j) = \frac{1}{(1 + d_j)} \tag{1}$$

The probability (1) helps in converting distances to a number between 0 and 1. Additionally, as the value of d_j increases, the probability decays very fast.

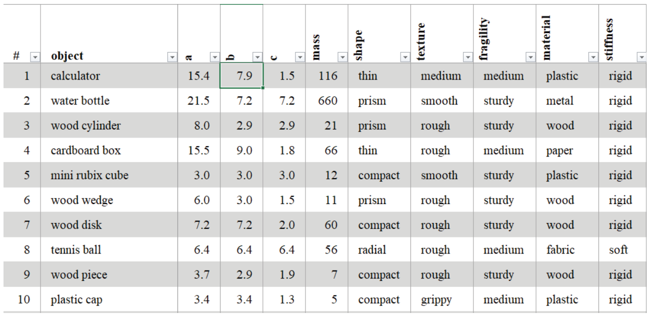
The overall probability of mapping i^{th} object to the reference object being described, is the joint probability over all the features of the object, given by:

$$P(i \rightarrow o|o) = \prod_{j=1}^{m} P(i \rightarrow o|d_j)$$

This approach ensures that only that object which matches the reference object’s every individual, available features is the one that results in a highest probability value.

TABLE II

OBJECT FEATURES DATASET – SAMPLE OF 10 OBJECTS



#	object	a	b	c	mass	shape	texture	fragility	material	stiffness
1	calculator	15.4	7.9	1.5	116	thin	medium	medium	plastic	rigid
2	water bottle	21.5	7.2	7.2	660	prism	smooth	sturdy	metal	rigid
3	wood cylinder	8.0	2.9	2.9	21	prism	rough	sturdy	wood	rigid
4	cardboard box	15.5	9.0	1.8	66	thin	rough	medium	paper	rigid
5	mini rubix cube	3.0	3.0	3.0	12	compact	smooth	sturdy	plastic	rigid
6	wood wedge	6.0	3.0	1.5	11	prism	rough	sturdy	wood	rigid
7	wood disk	7.2	7.2	2.0	60	compact	rough	sturdy	wood	rigid
8	tennis ball	6.4	6.4	6.4	56	radial	rough	medium	fabric	soft
9	wood piece	3.7	2.9	1.9	7	compact	rough	sturdy	wood	rigid
10	plastic cap	3.4	3.4	1.3	5	compact	grippy	medium	plastic	rigid

We then iterate over all the objects in the dataset and assign the respective joint probabilities of each of them being mapped to the reference object and would like to choose the one with the maximum joint probability over all the objects in the dataset. The object selected as the closest match to the reference object is given as

$$f^* = \arg\max_i \prod_{j=1}^{m} P(i \to o|d_j)$$

The joint probabilities tend to get numerically smaller. The equation can be log transformed to avoid numerical issues. Substituting (1) and applying transformation

$$f^* = \arg\max_i \sum_{j=1}^{m} \log_e \frac{1}{(1+d_j)}$$

Log probabilities are negative, so we take a negative of the log and look at the minimum over all observations. It follows that

$$f^* = \arg\min_i \sum_{j=1}^{m} \log_e (1 + |f_j(i) - f_j(o)|)$$

Let us call this the Joint Probability distance metric. Using this distance metric, the probability of the match decays at a much faster rate as each of the features deviate from the reference. That helps in nonlinearly increasing the distance of unlikely candidates and filtering out the unlikely matches with more confidence. This distance measure can be used on a dataset which contains a combination of continuous and categorical (transformed to one-hot binary encoding) features without any need for data normalization.

In this approach, f^* corresponds to an object in the dataset that results in the smallest distance between its features and that of the reference object. A feature of this approach is that, even when there is no exact match for the reference object in the dataset, the approach always finds and reports the closest available object, resulting in false-positive identifications of the object. This behavior, while not desirable, will still serve the ultimate objective of identifying a suitable grasp for the reference object. By identifying an object closely matching the description of the reference object, we continue to retain the ability of choosing the most suitable grasp because the reference object has physical features very similar to the object chosen by the algorithm.

Our implementation consisted of three steps as discussed below. The dataset was stored in comma separated text files and Python programming language with Sci-kit library [10] was used to perform the analysis. In summary, we used Python with NLTK library to parse the descriptions and to extract numerical data such as dimensions and mass. We score the extracted values against the actual values to validate the accuracy of the parsing algorithm. We use natural-language parsing to convert unstructured human description to a structured format. Using the structured form, we perform a distance based contextual search on the objects database and get the full feature definition of the object using the 'memory recall' method. To validate this 'memory recall' methodology, we used the natural language descriptions and parsed them into a structured list of features. This list of features extracted, usually contains only a subset of the features and they tend to deviate from the actual due to approximation errors which are common in human language. We use this subset features to query and match the object being described to an object in the database by using the Joint Probability distance metric. The accuracy of identification depends on the clarity, completeness, and precision of the natural language description. The more information the description contains, the more confidence there is in the object identification process.

IV. EXPERIMENTS

A. Data Collection

The data collection phase consisted of building two datasets, an object features dataset and a corresponding natural language descriptions dataset [12]. The datasets were created for the study of grasping of an anthropomorphic robot hand. Some samples of object features and natural-language descriptions are given in Table II.

1) Object Dataset: The guiding principal during the creation of the objects database was to ensure that we create a dataset that captures sufficient variation in objects features to capture as many variations in grasp choices as possible. A total of 100 objects of everyday use were compiled to create the object dataset that consists of dimensions, mass, material, texture, shape classifications.

Given the ergonomics of human grasps, it was expected to see the object sizes, shapes and mass to adhere to certain preferred range of values. The analysis of the collected data in Fig 3 shows that there is sufficient variation in the dataset compiled.

2) Natural Language Descriptions Dataset: Natural language descriptions of all the objects in the dataset were generated manually. The dataset was randomly split in half and assigned to two individuals. They were provided with a weighing scale and a ruler and asked to write down the descriptions of the objects including mass, shape, size, texture and rigidity.

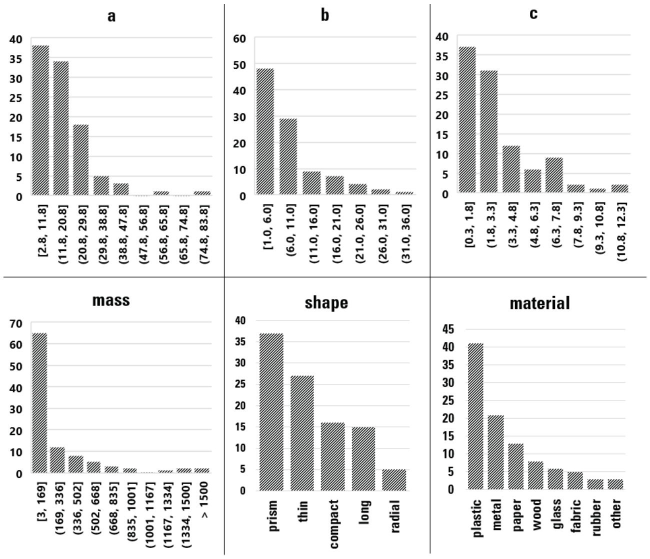
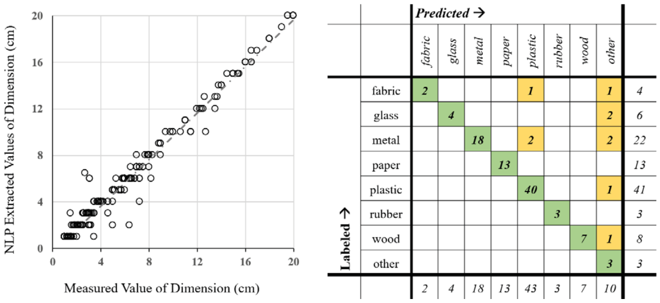


Fig. 3

Pareto histogram analysis of object features. The x axis is the value range of each feature and y axis represent number of object in that value range. Feature a, b and c are orthogonal dimensions of a object in cm, $mass$ is the weight of a object in $gram$.



	fabric	glass	metal	paper	plastic	rubber	wood	other	
fabric	2				1			1	4
glass		4						2	6
metal			18		2			2	22
paper				13					13
plastic					40			1	41
rubber						3			3
wood							7	1	8
other								3	3
	2	4	18	13	43	3	7	10	

Fig. 4

Left: Parsed vs measured dimensions. The closer the point is to the diagonal, the more accurate the NLP extract value is. Right: 'Material' type prediction score

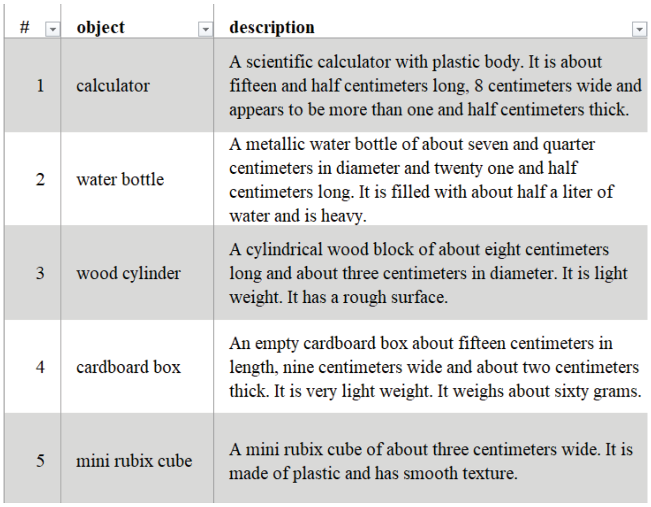
B. Natural Language Parsing

The effectiveness of the NLP algorithm was evaluated by scoring the output of the NLP with the actual values. A regression fit of the parsed vs actual values of dimension resulted in an R^2 of 0.98 (Fig 4) and for mass estimations the R^2 was 0.87. Categorical labels for material, shape and rigidity were scored as well. Fig 4 shows the scoring matrix for 'material'.

More detailed discussion regarding this NLP approach is dealt with in [12]. A quick take on the results is that the major error sources are: inaccurate or incomplete descriptions provided by humans combined with limitations of the parsing algorithm.

TABLE III

Object Descriptions Dataset – Sample Of 5 Objects



#	object	description
1	calculator	A scientific calculator with plastic body. It is about fifteen and half centimeters long, 8 centimeters wide and appears to be more than one and half centimeters thick.
2	water bottle	A metallic water bottle of about seven and quarter centimeters in diameter and twenty one and half centimeters long. It is filled with about half a liter of water and is heavy.
3	wood cylinder	A cylindrical wood block of about eight centimeters long and about three centimeters in diameter. It is light weight. It has a rough surface.
4	cardboard box	An empty cardboard box about fifteen centimeters in length, nine centimeters wide and about two centimeters thick. It is very light weight. It weighs about sixty grams.
5	mini rubix cube	A mini rubix cube of about three centimeters wide. It is made of plastic and has smooth texture.

C. Object Recall Accuracy

We started with only the natural-language description of each object and used the Joint Probability distance to match the reference object from the list of 100 objects. The accuracy was 92% i.e. 92 times out of 100, the natural language input provided returned the correct object type. The remaining 8 times, a different object was matched. Upon investigation, the sources of confusion were found to be:

- ◇ In 4 out of 8 instances, the descriptions were clearly incomplete. With insufficient information, there were multiple candidates that matched the descriptions and the algorithm chose one of them.
- ◇ In 3 out of 8 instances, the objects were very similar, and confusion was quite understandable. For example, a tennis ball and a plastic ball of similar sizes. The material description that could have been tie-breaker, was missing in the description and hence the confusion.
- ◇ In 1 instance, the rounded dimensions were quite off from the actual dimensions, stemming from the errors in parsing process and that resulted in the confusion.

D. Relative accuracy of various distance metrics

We repeated the previous step using the same set of inputs but different distance metrics namely Euclidean, Minkowski, Cosine and k-d Tree. We then validated the relative accuracies of each of these metrics in the context of the object recall problem. As can be seen from the Table 5, the proposed Joint Probability distance provides the means for most accurate recall of the object.

Distance Metric		Recall Accuracy	
Joint Probability	$\mathbf{d_{JP}} = \sum_{j=1}^{m} \log_e(1 + \lvert f_j^{(i)} - f_j^{(o)} \rvert)$	**92**/100	🟩 92 🟥 8
Euclidean	$\mathbf{d_{euc}} = \sqrt{\sum_{j=1}^{m} \lvert f_j(i) - f_j(o) \rvert^2}$	85/100	🟩 85 🟥 15
Minkowski p=4	$\mathbf{d_{Mk}} = \sqrt[p]{\sum_{j=1}^{m} \lvert f_j(i) - f_j(o) \rvert^p}$	82/100	🟩 82 🟥 18
Cosine	$\mathbf{S_{cos}} = \frac{\sum_{j=1}^{m} f_j(i) \cdot f_j(o)}{\sum_{j=1}^{m} f_j(i)^2 * \sum_{j=1}^{m} f_j(o)^2}$	59/100	🟩 59 🟥 41
K-D Tree	Algorithm	85/100	🟩 85 🟥 15

Fig. 5

COMPARISON OF OBJECT RECALL ACCURACIES OF VARIOUS DISTANCE METRICS.

E. Robot Simulation

In order to test the adequacy of this algorithm for robot applications, we fed the predicted output (matched objects) as input to the grasp planning module from [12]. The Python modules of this study were interfaced with grasp planning Python module which in turn was interfaced with ROS controlling the simulated AR10 Robotic Hand. Several different descriptions were tested. Two cases are discussed below.

Firstly, a description of a TV remote is used to test the algorithm. This specific remote was part of the database and it was made of plastic and had a long shape with a dimension of $16.0 \times 3.5 \times 2.0\, cm$. The intent was to check if the correct object would be identified. The algorithm matched correctly with the exact same TV remote in the database (item number 46). Since the object was correctly identified, there was no scope for confusion is grasp planning. The grasp type chosen was a Precision-Prismatic grasp[2] which is a feasible grasp type for the remote.

The second case was a description of an orange of 7 cm diameter. Here intentionally, the object was described which was not part of the database. The algorithm matched it with a Tennis Ball of 6.5 cm diameter in the database. The resulting grasp type chosen was a Power-Circular grasp [2] which is a feasible grasp type for both the ball and the orange.

In developing an algorithm to predict human grasp types using machine learning classifiers, it is important that we provide a reasonably accurate and structured representation of the object's features. This approach and the algorithm suggested in our study appears to be effective in object identification under uncertainty. The algorithm failed only in instances where the information available in the natural language description was incorrect or insufficient. Despite the confusion, it was interesting to note that the objects chosen by the algorithm were physically identical to the objects being described. The ultimate purpose of identifying the object is to choose an appropriate grasp. To that end, a mismatch here is not necessarily detrimental to the grasp choice. It is still possible to identify the correct grasp with the wrong object as long as the object is physically similar to the object being described.

V. CONCLUSION

Simply describing an objects features does not give one the ability to grasp it neither in humans nor in robots. It comes from referencing the information available with experience or familiarity of having held such objects. Recalling and knowing the specific object being described is an important skill that robots need to acquire before attempting a grasp. This study proposed a novel approach which uses a specific distance metric to identify the object being described. In the event of confusion, it still identifies the closest object with similar physical features, ensuring that there is enough information to continue onto grasp execution. Future explorations can be focused on modifying the algorithm and using large online datasets to expand the scope and improve the accuracy further. For now, the output of this study provides adequate information to predict certain types of grasps for humanoid robotic hands.

REFERENCES

[1] B. Burchfiel and G. Konidaris. Bayesian eigenobjects: A unified framework for 3d robot perception. In *Robotics: Science and Systems 2017*, 07 2017. 1

[2] M. R. Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on Robotics and Automation*, 5(3):269–279, Jun 1989. 2, 6

[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*, 46(1):66–77, Feb 2016. 2

[4] F. Heinemann, S. Puhlmann, C. Eppner, J. ÃL'lvarez Ruiz, M. Maertens, and O. Brock. A taxonomy of human grasping behavior suitable for transfer to robotic hands. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4286–4291, May 2015. 2

[5] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, J. Bohg, T. Asfour, and S. Schaal. Learning of grasp selection based on shape-templates. *Autonomous Robots*, 36(1):51–65, Jan 2014. 1

[6] R. S. Johansson and J. R. Flanagan. Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nature Reviews Neuroscience*, 10(5):345 – 359, 2009. 1

[7] S. Levine, P. Pastor, A. Krizhevsky, and D. Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *CoRR*, abs/1603.02199, 2016. 1

[8] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *CoRR*, abs/1703.09312, 2017. 1

[9] J. Napier. The prehensile movements of the human hand. *The Bone and Joint Journal*, 38 B(4), Nov 1956. 2

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 4

[11] L. Pinto and A. Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. *CoRR*, abs/1509.06825, 2015. 1

[12] A. B. Rao, K. Krishnan, and H. He. Learning robotic grasping based on natural language object descriptions. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018. 3, 4, 5, 6