Knowledge Augmentation and Task Planning in Large Language Models for Dexterous Grasping

Hui Li, Dang Tran, Xinyu Zhang and Hongsheng He*, *Senior Member, IEEE*

2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids) | 979-8-3503-0327-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/HUMANOIDS57100.2023.10375176

***Abstract*— Dexterous grasping is a critical ability for humanoid robots to interact efficiently with the physical environment. Human beings achieve dexterous grasping through a series of high level cognitive processes including target perception, object recognition, feature estimation, and intuitive reasoning. These processes cooperatively contribute to object understanding and the generation of appropriate grasping strategies. However, the current research focuses on establishing large object datasets to estimate object features and employing learning and planning approaches for task deployment, the exploration of the cognitive aspect of dexterous grasping is limited, especially the role of intuition. This paper addresses this research gap by investigating the cognitive processes in dexterous grasping and presents a cognition based grasping system. The proposed system integrates various cognitive processes to enable dexterous grasping. It gathers object information and estimates missing details using a large language model with common sense. Based on the complemented information, the system learns suitable grasp strategies and intuitively guides their execution. Real-world experiments with a anthropomorphic robot hand demonstrated the performance of the proposed system. By leveraging cognitive processes and utilizing the capabilities of a large language model, The proposed method enhances object understanding, generates effective grasping strategies, and provides guidance for the execution of the grasping strategies.**

I. Introduction

In recent years, the need for humanoid robots is consistently growing in our daily lives such as home service, education, medication, and transportation [1]–[4]. The robots are required to interact with real-world environments with efficiency and dexterity, making dexterous manipulation the most challenging problem [5]–[7]. Dexterous grasping, as the initial step of dexterous manipulation, mainly relies on a comprehensive understanding and precise measurement of the physical characteristics of the target object, including dimensions, material, rigidity, texture, fragility, and shape, which influence grasping strategies significantly [8], [9]. Although many methods have been developed to measure the physical features of objects [10], [11], humanoid robots often struggle to gather complete information due to factors such as lighting conditions, obstacles, and varying viewpoints. Consequently, it is therefore beneficial to research an approach to plan dexterous grasping without complete or accurate sensing of object characteristics.

To address this problem, knowledge datasets of objects were established either by collecting the features manually or by mining the data from the internet [12]–[14]. However, building such large datasets requires enormous resources and can be impractical. Some research try to bypass the problem by learning directly from human demonstrations [14]. However, human grasping also needs common sense to estimate contextual information and decide grasping strategies. Inspired by human grasping, we address this problem by employing large language models (LLM). LLMs, such as GPT-3, BERT, XLNet, and T5, are trained with extensive text data to learn patterns, relationships, and context. They possess a basic understanding of objects and some level of common sense, which are leveraged to complement the missing information about the target object [15].

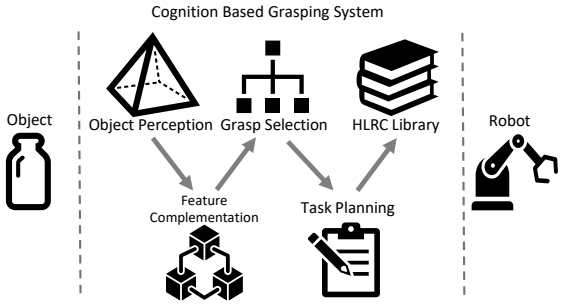

Fig. 1. Framework of the cognition based grasping system.

Human grasping is a complex process. Dexterity in grasping and manipulation is enabled by the redundant DoF of multi-fingered hands. There may be many possible strategies to grasp an object, and the optimal one depends on the affordances of the target object. Previous works have successfully mapped object features to proper grasp strategies [16], [17], yet the deployment of these strategies remains a challenging task. The current research employs learning or planning approaches to address this problem, but these methods either require a large amount of data or a comprehensive understanding of the grasping environment. In this work, we further explore the intuitive reasoning abilities of the large language model to tackle this problem efficiently.

In this paper, we propose a cognition based grasping system, as shown in Fig. 1, to guide humanoid robots in accomplishing dexterous grasping of daily objects using incomplete features. The system consists of five models: Object Perception (OP), Feature Complementation (FC), Grasp Selection (GS), Task Planning (TP), and High Level Robot Control Library (HLRC). The workflow of the

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

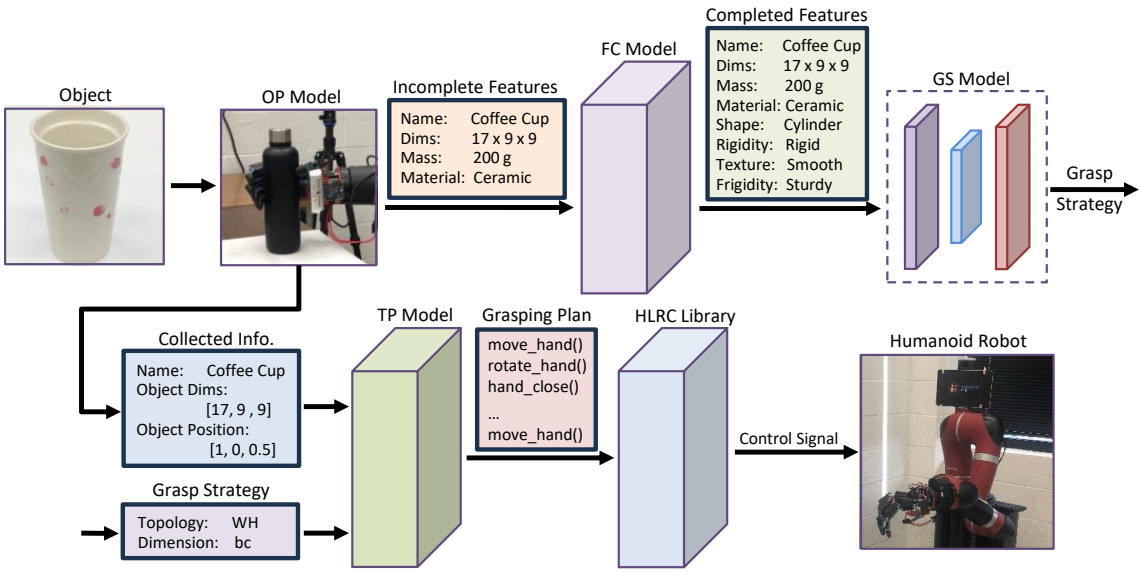

Fig. 2. The workflow of the proposed system. In the figure, OP stands for Object Perception, FC represents Feature Complementation, GS denotes Grasp Selection, TP is the abbreviation for Task Planning, and HLRC is short for High Level Robot Control Library.

system is illustrated in Figure 2. The Object Perception model perceives the object and gathers useful information about both the object and the environment. The Feature Complementation model estimates and complements the features collected in the Object Perception model. The Grasp Selection model maps the complemented features to appropriate grasp strategies. The Task Planning model generates functional code to guide the robot in executing the task, and the High Level Robot Control Library provides functions to adapt to the code and control the robot. The contributions of this research are twofold:

- ◇ Design and implement a cognition based grasping system that can tackle grasping task dexterously using incomplete information of the target object.
- ◇ Conducted real-world experiments using a humanoid robot hand to demonstrate the performance of the proposed system.

II. Cognition Based Grasping System

Dexterous grasping in unstructured environments could be challenging due to the lack of environmental information and the ability for intuitive reasoning and planning. We, therefore, developed a cognition based grasping system to solve this problem. This system complements perceived object features, maps the features to grasp strategies, and guides and performs the grasping task using common sense and intuitive reasoning. In this section, we provide a detailed explanation of the proposed system.

A. Object Perception

The Object Perception model utilizes the MagicHand platform to acquire object features, including object name, mass, material, and dimensions, and information about the environment such as the position and orientation of the object. The MagicHand platform integrates multiple sensors, including RGB-D cameras, Force Sensing Resistors (FSR), and a SCiO sensor, to enhance perception abilities. RGB-D cameras are used for object recognition, 3D modeling, and determination of object coordinates and orientation. FSR sensors are integrated into the tips of a five-digit robotic hand to enable the detection of contact force. The SCiO sensor collects the near-infrared spectrum of the object for material recognition [18]. An electrical scale is also employed in the model to measure the mass of the object. The dimensions of the target are derived from its 3D model. Further details of the MagicHand platform can be found in our previous work [19].

B. Feature Complementation

Although certain information about the object and the environment can be collected by the Object Perception model, perceiving features such as fragility, rigidity, shape, and textures can still be challenging. Traditional methods for acquiring these attributes, often involve costly chemistry or physical analysis and are not suitable for real-time dexterous grasping. It is straightforward to estimate these attributes based on the recognized features using common sense, similar to human reasoning. We, therefore, developed a Feature Complementation model by leveraging a large language model which has a basic understanding of common objects and common sense.

Even though large language models possess knowledge and proficiency in common-sense and logical reasoning, they demonstrate limitations in problem-solving [20]. LLMs are sensitive to input phrasing and often output misleading and overgeneralized knowledge. To address this issue, we precisely define and categorize each feature and build

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

prompts to prompt the LLM to generate consistent and unambiguous estimations of missing object features.

1) Feature Definition and Categorization: We provide clear definitions and categorizations for physical characteristics including fragility, rigidity, shape, and texture. For each feature, we explored different definitions of that feature from various dictionaries, online resources, and research papers. Then, we input each definition into the LLM and observed the output feature name. The definition that yields the most consistent and accurate results was selected as the definition for that particular feature.

While estimating object features, humans often provide a binary response, such as "smooth" or "tough" for texture and "rigid" or "soft" for rigidity. Thus, we emulate this strategy to categorize features including fragility, rigidity, and texture. However, classifying the shape of an object into binary classes is not possible. Instead, we categorize the shape with the most basic shapes in geometry. In addition, a complex shape can be treated as a combination of these basic shapes. The definitions and categorizations for each feature are shown in Table I.

TABLE I
FEATURE DEFINITIONS AND CATEGORIZATIONS.

Features	Definition	Categorizations
Fragility	Tendency to break, shatter, or deform when subjected to external forces or stress.	{fragile, sturdy}
Rigidity	Ability to resist deformation or bending and retain its shape and structural integrity when external forces are applied.	{soft, rigid}
Shape	The overall form and structure of an object.	{sphere, cube, cone, cylinder, cuboid, disk}
Texture	physical characteristics and qualities, such as roughness or smoothness, of the outer layer of an object	{smooth, rough}

These definitions and categorizations reduce ambiguity and enhance the LLM's comprehension of those features, thus achieving a more accurate understanding of the attributes.

2) Prompt Implementation: The accuracy and the consistency of LLMs highly depend on the quality of the descriptions of the problem such as phrasing, relevant details, context, and specific requirements or constraints. To have LLMs generating more close output to the desired one, we formulated various prompts to enhance problem descriptions. The prompt yields the most precise and consistent result is shown in Table II. An example input "calculator 15.4 7.9 1.5 116 plastic" yields the result "{"shape": "cuboid", "texture": "smooth", "rigidity": "rigid", "fragility", "sturdy"}".

TABLE II
PROMPT FOR FEATURE COMPLEMENTATION.

Imagine you are helping me to estimate the physical features of an object based on some known features of that object. I will give you some features of an object, and you need to complement the features with common sense. The features that need to be complemented are Fragility, Shape, Texture, and Rigidity. Shape is defined as the overall form and structure of an object and is categorized as {sphere, cube, cylinder, cone, cuboid, disk}. Texture is defined as the physical characteristics and qualities, such as roughness or smoothness, of the outer layer of an object, and categorized as {smooth, rough}. Rigidity is defined as an object's ability to resist deformation or bending and retain its shape and structural integrity when external forces are applied, and categorized as {soft, rigid}. Fragility is defined as an object's tendency to break, shatter, or deform when subjected to external forces or stress, and categorized as {fragile, sturdy}. The output should be in JSON format. Only output Fragility, Shape, Texture, and Rigidity. Do not explain your answer.

C. Grasp Selection

The grasp of an anthropomorphic robotic hand defines a set of angles of the finger joints, and the magnitude of the contact forces applied by the fingers and palm to an object at the contact points. The objective here is to emulate human grasping by mapping object features f complemented from the Feature Complementation model onto grasp prioritization H

$$f = [a, b, c, m, s, mt, r, t, fr] \rightarrow H \tag{1}$$

where $(a,\ b,\ c)$ are dimensions along orthogonal directions $(a \geqq b \geqq c)$ and $(m,\ s,\ mt,\ r,\ t,\ fr)$ are the mass, shape, material type, rigidity, texture and fragility of the object.

1) Grasp Definition: The complemented object features f may be inaccurate or even erroneous due to rough measurement and estimation. To enable imprecise measure of object dimensions and position, and improving system robustness and adaptivity, We therefore implement grasp strategies in terms of grasp topology and grasp dimension. Grasp scales are determined by the orthogonal dimensions a, b, c around which the grasp closure occurs, as illustrated in Fig. 3. This labeling convention is commonly adopted in the literature [21], facilitating the computation of hand closure in forward and inverse kinematics. Grasp dimension d includes all feasible dimensions that can be utilized to grasp the object and

$$d \in \{a, b, c, ab, bc, ac, abc\} \tag{2}$$

where ab indicates the object can be grasped either around dimension a or around dimension b.

The grasp topology h is one of the grasp types drawn from the set of human grasp primitives

$$h \in \{wt, wp, wh, wc, rp, rc\} \tag{3}$$

where grasp types wt, wp, wh, wc, rp and rc, as shown in Figure 4, are high level grasp topology adopted from the

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

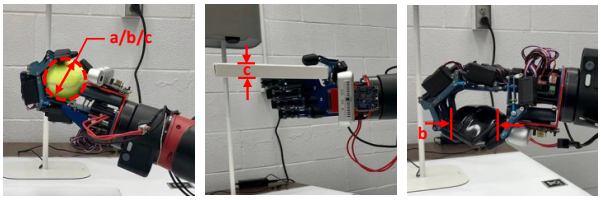

Fig. 3. Illustrations of grasp dimension.

grasp taxonomy presented by Cutkosky [22]. This grasp definition can be easily applied to other type of hand-effectors by restricting the number of fingers used.

We developed a grasp classification system by combining the grasp type and dimension to overcome the inconsistencies in grasp dimensions for certain object-grasp associations. This taxonomy ensures that each grasp topology is associated with proper dimensions and sizes. The extended grasp taxonomy is defined as

$$H \in \{rc.ab, rc.bc, rp.b, rp.c, wc.abc, wh.bc, wh.c, wp.bc, wt.c\} \tag{4}$$

Further details regarding this methodology can be found in our previous work [17].

2) Learning Grasping Strategies from Object Features: Most studies attempting to understand and codify human grasps have concluded that human grasp choice is a function of object affordances (geometry, texture etc.) and the task requirements (forces, mobility, etc.) [8], [22], [23]. Attempts to assign one most suitable grasp for a given object-task combination have not been conclusive. The major problem is that even for the same specific object-task combination, there are multiple grasp choices possible, which appears to be arbitrary and not amenable for deterministic modeling. Human grasp choices nevertheless do tend to cluster when studied over a large set of objects. Both the clustering effect and the confusion between grasp types can be seen in the data presented by [21], which shows that a single object could be held in multiple different grasp types in the course of picking or handling. There is no one-to-one mapping of one object to one grasp type.

The problem of grasp selection is therefore not selecting one ideal grasp type but one of the many feasible grasp types in human grasp taxonomy for the given context. To that end, we plan to learn the mapping from features f into grasp topology distributions

$$f \rightarrow P(H|f) = [P(rc.ab|f), \cdots, P(wt.c|f)] \tag{5}$$

We designed a neural network to model the probability distribution over all grasp classes $\hat{P}(H|f)$, as illustrated in Fig. 5. The network is designed with cross-entropy loss and optimized using stochastic gradient descent algorithms. The loss function is defined by cross entropy that measures the deviation between the ground truth and predicted probability distribution

$$L(P(H|f), \hat{P}(H|f)) = \sum_i \sum_j P(H_j|f_i) \log \hat{P}(H_j|f_i) \tag{6}$$

where $i \in [1, N]$ with N as the number of observations and j is the index of grasp topology.

The grasp with the maximum probability is chosen by

$$\hat{H}^*_{\max} = \arg\max_j \hat{P}(H|f) \tag{7}$$

the predicted grasp configuration $\hat{H}^*_{\max}$ contains information regarding the grasp type and object dimension along which the grasp can be executed, so $\hat{H^*}_{max}$ can be easily decomposed into grasp type h^* and grasp dimension d^*, which can be used subsequently to calculate robot hand configuration. The optimal grasp type is chosen as the one corresponding to the highest probability from the predicted probability distribution.

Because the model predicts probability distributions, we defined two scoring metrics for training and evaluation of the model. The predicted grasp choice is scored as a success if the same grasp type was chosen at least once in the evaluation dataset. The feasibility of the grasp is scored as

$$F_l(P(H), \hat{P}(H)) = \begin{cases} 1 & P(\hat{H}_{\max}) > 0 \\ 0 & P(\hat{H}_{\max}) = 0 \end{cases} \tag{8}$$

where

$$\hat{H}_{\max} = \arg\max_j \hat{P}(H_j|f) \tag{9}$$

is the grasp topology with the maximal probability, and H is defined in (4). The feasibility score F_l is representative of the ability of the algorithm to pick a feasible grasp for a given object. The match score metric F_m is defined as

$$F_m(P(H), \hat{P}(H)) = \begin{cases} 1 & P(\hat{H}_{\max}) = P(H_{\max}) \\ 0 & P(\hat{H}_{\max}) \neq P(H_{\max}) \end{cases} \tag{10}$$

This match score is representative of the ability of algorithm to predict the most frequently applied human grasp as the grasp with the highest probability for a given object. In other words, F_m is akin to the accuracy. This metric is much more stringent and therefore we can expect the match score F_m to be always lower than feasibility score F_l

$$F_m(P(H), \hat{P}(H)) \leq F_l(P(H), \hat{P}(H)) \tag{11}$$

We used the feasibility score as the primary scoring metric, for the objective is to find one feasible grasp that can be successfully executed by a robot.

D. Task Planing

Traditional grasp deployment methods, relying on planning or learning approaches, require a deep understanding of the environment or extensive data for training which are not resource-efficient and sometimes impossible to achieve. In this section, we introduce a LLM-based Task Planing model, which leverages LLM's

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

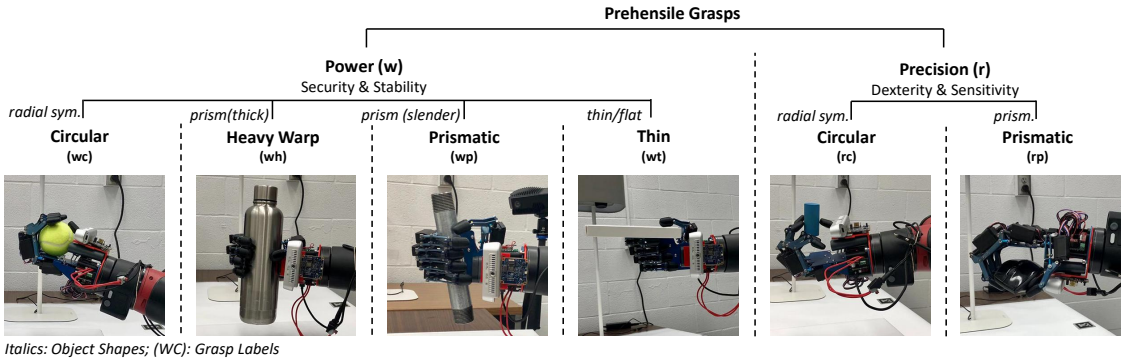

Fig. 4. Human grasp taxonomy derived from [22].

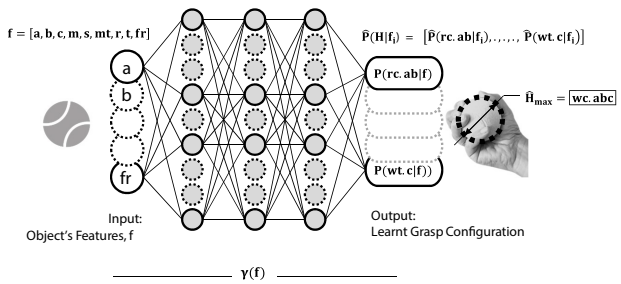

Fig. 5. The Grasp Selection model.

abilities of common sense and reasoning abilities, to deploy the learned grasp strategies with efficiency and flexibility.

```
Python
# Step 1: Decide the most secure and stable grasp direction from the sides
grasp_direction = 'side'
# Step 2: Rotate hand and approach the object based on the selected
direction
rotate_hand(grasp_direction)
approach_object(grasp_direction)
# Step 3: Move to the object
object_position = get_object_position(grasp_direction)
target_position = [ object_position[0], object_position[1],
object_position[2]]
move_hand(target_position)
# Step 4: Grasp the object
grasp_object()
# Step 5: Pick up the object for 10 cm
target_position[2] += 10 # Move 10 cm upward from the current position
move_hand(target_position)
```

Fig. 6. Generated grasping deployment plan.

To generate appropriate grasping plans, we devised prompts that outline the task, elucidate the available functions, provide a description of the environment, and specify the goal. The model will analyze the given information, decide to grasp the object from the top or from the side, and generate a comprehensive and practical grasping deployment plan based on the decision. To simplify the problem, grasp pose is pre-adopted and the target object was placed in a manner that its grasp dimension can be grasped from the direction decided by the model. The final prompt,

TABLE III
PROMPT FOR TASK PLANING.

An object is placed on a table in front of the robot. The the width of the palm is 11 cm. The max length of the grasp aperture is 12 cm. The height of the object is along the vertical direction, width along horizontal direction, and the thickness along the forward-backward direction. Your task is to grasp the object and pick it up. To accomplish the task, you need to 1. decide which direction (side or top) would provide the most secure and stable grasp. 2. rotate hand and approach the object based on the selected direction. 3. move to the object. 4. grasp the object. 5. pick up for 10 cm. The following functions are available for you: rotate_hand(direction): take a string input, rotate the robotic hand so that the hand can grasp the given direction approach_object(direction): move to a safe distance based on the grasp direction get_object_position(direction): return the current coordinates and orientation of the direction of the object. move_hand([x,y,x]): move the hand to the given position [x,y,x] grasp_object(): close the hand to grasp the object. You need to learn the skill of picking up the object and holding it.The object is a water bottle with dimension $21 \times 7.5 \times 7.5$ corresponding the height, the width, and the thickness. Use your common sense and reasoning skill to write python code to control the robot to pick up this specific object. You are allowed to create new functions using the available functions, but you are not allowed to use any other hypothetical functions. Keep the solutions simple and clear. Do not output code for each step. Output an overall code. Additional points to consider when giving your answer: 1. Your responses should be informative, visual, logical and actionable. 2. Your logic and reasoning should be rigorous, intelligent, and defensible. 3. You can provide additional relevant details to respond thoroughly and comprehensively to cover multiple aspects in depth.

shown in Table III, yields the grasping deployment code in Figure 6. Note that the prompt is task and environment specific. For different task and environment combinations, the prompt needs to be adjusted accordingly.

E. High Level Robot Control Library

This High Level Robot Control Library (HLRC) is implemented to adapt to the code generated from the Task Planing model. HLRC is a versatile library developed, utilizing API from OpenCV, Sawyer and AR10 libraries, and ROS Python libraries, to provide precise robot control

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

in real-world robot test beds. The "get_object_position" function locates the object and return its current coordinates and orientations in the environment through the MagicHand system [24]. The "rotate_hand" function rotates the robot hand so that the robot can grasp the object from either the left side or the top. The "approach_object" function ensures the object in the grip by calculating the relative position among the palm, the fingers and the object. The "grasp_object" function executes the grasping action until a specific contact force threshold is reached.

III. Experiments

Real-world experiments were conducted using an anthropomorphic robotic hand to evaluate the overall performance of the proposed cognition based grasping system. We revised our previous object dataset in [13] to test the accuracy of the Feature Complementation model and to train and evaluate the Grasp Selection network. The experiment results demonstrated the performance and improved the overall understanding of the proposed system.

A. Experiment Setup

1) Testing Environment: The MagicHand platform, as shown in Fig. 7, serves as the test bed for the proposed system. The robot includes an AR10 robotic hand, a Rethink Sawyer robot (with an in-arm camera), as well as an Intel RealSense RGB-D camera, Force Sensing Resistors (FSR), and an SCiO sensor installed on the wrist. The robotic hand has limited force sensing through the FSR attached to fingertips, so we focused on hand configurations and planning and used the force sensors to examine contact conditions.

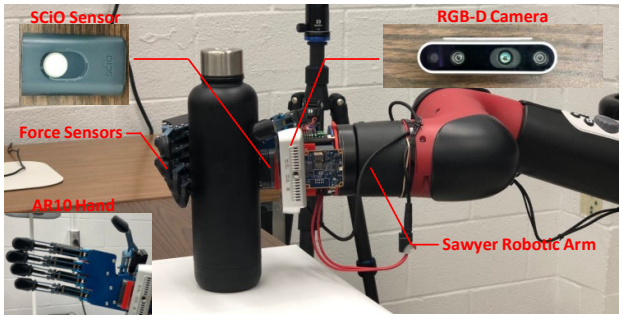

Fig. 7. MagicHand: Sawyer robotic arm with an AR10 robotic hand and in-hand sensors.

2) Testing Dataset: To align with the categorization strategy described in Section 2. B, we revise the Fragility, Rigidity, Shape, and Texture features of the object dataset [13]. The dataset was established by nine non-expert college students who collected information from 100 everyday objects and labeled each object with all applicable grasp topologies. Object features, including name, dimensions, mass, shape, texture, fragility, material, and rigidity, were measured, estimated, and mapped with grasp topologies in 4. For each object, the label includes all applicable grasp topologies and their corresponding frequency. A selection of object samples is show in Figure 8.

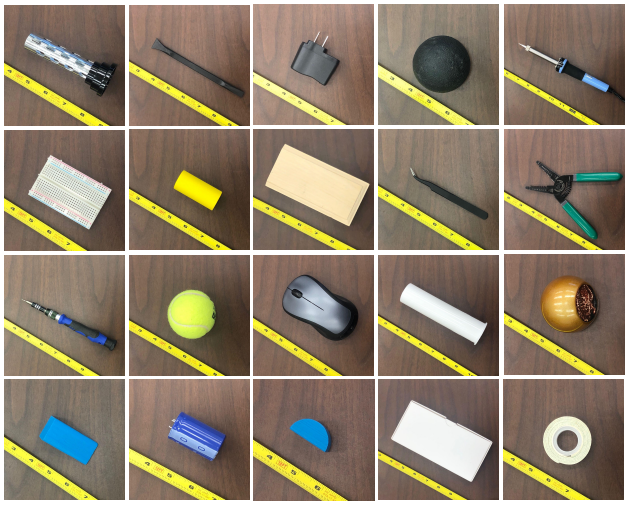

Fig. 8. Sample objects for experiments.

B. Object Feature Estimation

The efficiency of the Feature Complementation (FC) model is evaluated with the revised object dataset. Perceived information from the Feature Perception model, including object name, dimension, and material, is enhanced and complemented. The performance of the FC model is shown in Figure 9.

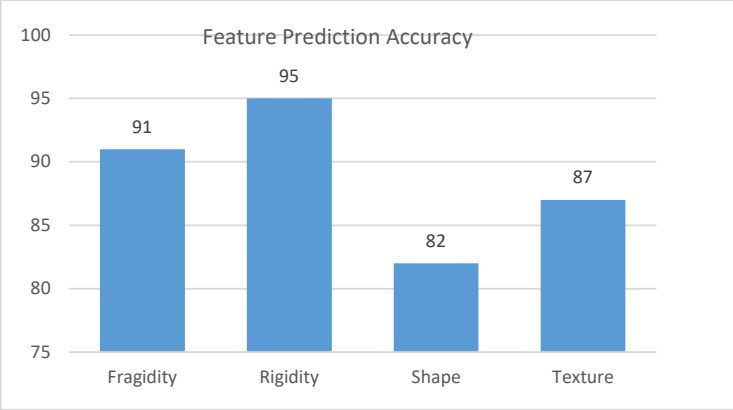

Fig. 9. Feature estimation accuracy of the Feature Complementation model.

In the figure, we can see that the model has a lower accuracy in predicting shape and texture. This is due to the limitation of the given information. Objects with the same names and dimensions could have different shapes. For example, a $7 \times 7 \times 2$ plate could either be square-shaped or disk-shaped. Additionally, differentiating cylinder and cuboid based on the dimensions of the object can be also challenging. By using diameter to describe the disk and cylinder-shaped object, the accuracy for predicting shape increased to 91%. The prediction of the texture faces the same problem, name and material of the object are not

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

	Labled Grasp Probability Distributions									Predicted Grasp Probability Distributions										
Object	rc.ab	rc.bc	rp.b	rp.c	wc.abc	wh.bc	wh.c	wp.bc	wt.c	rc.ab	rc.bc	rp.b	rp.c	wc.abc	wh.bc	wh.c	wp.bc	wt.c	Fl	Fm
calculator	0.00	0.00	0.56	0.22	0.00	0.00	0.00	0.00	0.22	0	0.22	0.41	0	0	0	0	0	0.37	1	1
water bottle	0.00	0.11	0.11	0.22	0.00	0.56	0.00	0.00	0.00	0	0	0.32	0.25	0	0.43	0	0	0	1	1
wood cylinder	0.00	0.22	0.56	0.22	0.00	0.00	0.00	0.00	0.00	0	0.18	0.31	0.29	0	0.22	0	0	0	1	1
cardboard box	0.00	0.00	0.56	0.22	0.00	0.00	0.00	0.00	0.22	0	0	0.37	0.19	0	0	0.22	0	0.22	1	1
mini rubix cube	0.11	0.44	0.33	0.11	0.00	0.00	0.00	0.00	0.00	0	0.45	0.31	0	0	0	0	0	0.24	1	1
wood wedge	0.00	0.22	0.44	0.22	0.00	0.00	0.00	0.00	0.11	0	0.12	0.25	0.43	0	0	0	0	0.2	1	0
wood disk	0.44	0.00	0.22	0.22	0.00	0.00	0.00	0.00	0.11	0.34	0	0.26	0.4	0	0	0	0	0	1	0
tennis ball	0.11	0.11	0.22	0.11	0.44	0.00	0.00	0.00	0.00	0.2	0.25	0	0.25	0.3	0	0	0	0	1	1
wood piece	0.22	0.44	0.22	0.11	0.00	0.00	0.00	0.00	0.00	0.11	0.39	0.26	0.24	0	0	0	0	0	1	1
plastic cap	0.56	0.00	0.22	0.11	0.00	0.00	0.00	0.00	0.11	0.15	0	0.25	0.2	0.2	0	0	0	0.2	1	0
medicine dispenser	0.00	0.00	0.67	0.22	0.00	0.00	0.00	0.11	0.00	0	0	0.35	0.18	0	0	0.22	0.25	0	1	1
screw driver	0.00	0.00	0.33	0.11	0.00	0.00	0.00	0.56	0.00	0	0	0.17	0.29	0	0.16	0	0.38	0	1	1
ball1	0.11	0.22	0.11	0.11	0.44	0.00	0.00	0.00	0.00	0.19	0.15	0.21	0	0.45	0	0	0	0	1	1
rubix cube	0.11	0.11	0.56	0.22	0.00	0.00	0.00	0.00	0.00	0.22	0.2	0.3	0.16	0.12	0	0	0	0	1	1
water bottle cap	0.56	0.00	0.33	0.11	0.00	0.00	0.00	0.00	0.00	0.35	0	0.23	0.23	0.19	0	0	0	0	1	1
sanitizer bottle	0.00	0.00	0.56	0.22	0.00	0.22	0.00	0.00	0.00	0	0.25	0.4	0.1	0	0.25	0	0	0	1	1
wallet	0.11	0.00	0.44	0.22	0.00	0.00	0.00	0.00	0.22	0.25	0	0.35	0.24	0	0	0	0	0.16	1	1
ball2	0.11	0.22	0.11	0.11	0.44	0.00	0.00	0.00	0.00	0.21	0.24	0	0	0.55	0	0	0	0	1	1
toy car	0.00	0.00	0.56	0.33	0.00	0.11	0.00	0.00	0.00	0	0	0.48	0.23	0	0.29	0	0	0	1	1
kitchen scale	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.67	0	0	0.18	0.27	0	0	0	0	0.55	1	1

Fig. 10. Sample grasping strategy determination: ground-truth vs. predicted grasping.

enough for deciding whether its surface is smooth or rough. The surface of objects can vary in smoothness or roughness not only based on the material but also due to the surface structure. For instance, the surface of a metal cup can be either smooth or rough based on the surface finish or polishing.

C. Grasp Selection Network

We developed the neural-network model to learn grasping strategies as proposed in Section 2. C, and optimized the model in terms of cross-entropy. The input layer includes nine nodes activated by ReLu function. The next four layers are hidden layers which contains 2^7, 2^9, 2^7 and 2^5neurons which are also activated by ReLu function. The output layers has nine nodes activated by Sigmoid activation function. The input of the model is the acquired object features, and the output is the grasping strategies corresponding to human preference and knowledge. The grasping strategies were represented by normalized probability distributions. The results of grasping strategy determination with scores are reported and compared to the ground truth in Fig. 10. The feasibility score F_l of the model is 100%, which was defined as the hit rate of the predicted grasp strategy in all human preferred grasps. The experiment shows that the model's capability in picking feasible (human validated) grasping. The match-score F_m, on the other hand, measures the accuracy of the prediction considering only the most preferred human grasping. The experiment demonstrated that the max-match rate was around 85% for the test objects. The objects with complemented features were also tested by the network. The testing result shows a feasibility score of 100% and a match-score of 83% which is similar to the object with human-decided features, indicating the grasp topology predicted by the grasp selection network based on complemented features are comparable to those made by human decision-makers.

D. Robotic Grasping

To further examine the performance of the proposed system, we performed grasping experiments using the MagicHand platform. Test objects were placed on the table with a default initial orientation, and located and identified by the Object Perception model. The perceived information was then complemented by the Feature Complementation model. The robot autonomously chose the grasp strategies according to the completed object affordance. The success of a grasp was determined by the stability of grasping after brief maneuvers including grasping, lifting, and holding. A grasping task was considered a success if the object stays secure throughout the maneuvers.

We conducted ten grasping tasks on different objects and the overall success rate of grasping was 80%. One failure case involved the grasping of a large-sized metal cup, where the model predicted the most preferred human grasping strategy (wh) and grasp direction (side). However, the hand failed to wrap the cup tightly and securely, leading to the failure of the grasping task. This was attributed to the limitations of hand dexterity and the usage of the pre-defined grasp pose. The other failure case happened while trying to grasp a small bottle cap. Even though the model predicted the most preferred human grasp (rp), it decided to grasp the object from the side, we terminated the task to prevent potential collision between the hand and the table.

We designed a sequential system to emulate human

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.

decision-making processes. In such a system, errors from one model could permeate or propagate to subsequent models. However, in the experiments, the final grasp execution on the robot was minimally affected by the errors generated in each model. The primary reason could be that, while human grasping is complex, it is also highly resilient to external perturbations of contextual variables. When modeling robot grasping using human grasp primitives, this resilience behavior was emulated as well. For example, there are multiple ways to grasp an object, so it is less likely to choose a wrong grasp as we have seen from the results of our Grasp Selection model. The other reason is that the experiment objects were designed with the intent of handling by five-fingered hands, so when there is a miscalculation, e.g., grasp dimensions, the fingers conform to the object shape and still result in a secure grasp.

IV. Conclusion

This paper has presented a cognition based grasping system to applying a proper strategy to grasp an object without complete sensing of object affordance. The framework of the proposed system emulates the human grasping process, including object affordance acquisition, strategy determination, and grasping deployment, by combining common sense, intuitive, reasoning, and machine learning approach. The accuracy of the feature complementation process are between 82% to 95% depending on the feature. The grasp selection model achieves a 100% feasibility score and an 85% match score in predicting human grasping knowledge. Experiment on the humanoid robot achieved 80% success rate, demonstrating the practicality and efficiency of the proposed system in dexterous grasping task in unfamiliar environments. In summary, the experiments show that the proposed system is efficient and suitable for real-world grasping applications.

References

[1] I. Papadopoulos, C. Koulouglioti, R. Lazzarino, and S. Ali, "Enablers and barriers to the implementation of socially assistive humanoid robots in health and social care: a systematic review," *BMJ open*, vol. 10, no. 1, p. e033096, 2020.

[2] Y. Choi, M. Choi, M. Oh, and S. Kim, "Service robots in hotels: understanding the service quality perceptions of human-robot interaction," *Journal of Hospitality Marketing & Management*, vol. 29, no. 6, pp. 613–635, 2020.

[3] I. Papadopoulos, R. Lazzarino, S. Miah, T. Weaver, B. Thomas, and C. Koulouglioti, "A systematic review of the literature regarding socially assistive robots in pre-tertiary education," *Computers & Education*, vol. 155, p. 103924, 2020.

[4] X. Yu, B. Li, W. He, Y. Feng, L. Cheng, and C. Silvestre, "Adaptive-constrained impedance control for human–robot co-transportation," *IEEE transactions on cybernetics*, vol. 52, no. 12, pp. 13237–13249, 2021.

[5] M. Li, K. Hang, D. Kragic, and A. Billard, "Dexterous grasping under shape uncertainty," *Robotics and Autonomous Systems*, vol. 75, pp. 352–364, 2016.

[6] F. Chen, M. Selvaggio, and D. G. Caldwell, "Dexterous grasping by manipulability selection for mobile manipulator with visual guidance," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 1202–1210, 2018.

[7] M. Indri, A. Grau, and M. Ruderman, "Guest editorial special section on recent trends and developments in industry 4.0 motivated robotic solutions," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1677–1680, 2018.

[8] 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.

[9] 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.

[10] G. N. DeSouza and A. C. Kak, "Vision for mobile robot navigation: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 2, pp. 237–267, 2002.

[11] R. Fernandez-Rojas, A. Perry, H. Singh, B. Campbell, S. Elsayed, R. Hunjet, and H. A. Abbass, "Contextual awareness in human-advanced-vehicle systems: A survey," *IEEE Access*, vol. 7, pp. 33304–33328, 2019.

[12] B. Rao, H. Li, K. Krishnan, E. Boldsaikhan, and H. He, "Knowledge-augmented dexterous grasping with incomplete sensing," *arXiv preprint arXiv:2011.08361*, 2020.

[13] 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)*, pp. 1368–1373, IEEE, 2019.

[14] C. Della Santina, V. Arapi, G. Averta, F. Damiani, G. Fiore, A. Settimi, M. G. Catalano, D. Bacciu, A. Bicchi, and M. Bianchi, "Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1533–1540, 2019.

[15] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *Microsoft Auton. Syst. Robot. Res*, vol. 2, p. 20, 2023.

[16] H. Li, Y. Zhang, Y. Li, and H. He, "Learning task-oriented dexterous grasping from human knowledge," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6192–6198, IEEE, 2021.

[17] 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)*, pp. 882–887, IEEE, 2018.

[18] H. Li, Y. Yihun, and H. He, "Magichand: In-hand perception of object characteristics for dexterous manipulation," in *Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10*, pp. 523–532, Springer, 2018.

[19] H. Li, J. Tan, and H. He, "Magichand: Context-aware dexterous grasping using an anthropomorphic robotic hand," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9895–9901, IEEE, 2020.

[20] N. Bian, X. Han, L. Sun, H. Lin, Y. Lu, and B. He, "Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models," *arXiv preprint arXiv:2303.16421*, 2023.

[21] 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, pp. 66–77, Feb 2016.

[22] M. R. Cutkosky, "On grasp choice, grasp models, and the design of hands for manufacturing tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, pp. 269–279, Jun 1989.

[23] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks.," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345 – 359, 2009.

[24] H. Li, J. Tan, and H. He, "Magichand: Context-aware dexterous grasping using an anthropomorphic robotic hand," in *IEEE International Conference on Robotics and Automation*, 2020.

Authorized licensed use limited to: UNIV OF ALABAMA-TUSCALOOSA. Downloaded on February 08,2024 at 13:29:42 UTC from IEEE Xplore. Restrictions apply.