IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Visual-Tactile Robot Grasping based on Human Skill Learning from Demonstrations using A Wearable Parallel Hand Exoskeleton

Zhenyu Lu[1]*, Lu Chen[2]*, *Member, IEEE*, Hengtai Dai[1]*, Haoran Li[3], Zhou Zhao[4], Bofang Zheng[1], Nathan F. Lepora[3], *Member, IEEE*, and Chenguang Yang[1†], *Senior Member, IEEE*

***Abstract*— The soft fingers and strategic grasping skills enable the human hands to grasp objects in a stable manner. This paper is to model human grasping skills and transfer the learned skills to robots to improve grasping quality and success rate. First, we designed a wearable tool-like parallel hand exoskeleton equipped with optical tactile sensors to acquire multimodal information, including hand positions and postures, the relative distance of the exoskeleton claws, and tactile images. Using the demonstration data, we summarized three characteristics observed from human demonstrations, involving varying-speed actions, grasping effect read from tactile images and grasping strategies for different positions. The characteristics were then utilized in the robot skill modelling to achieve a more human-like grasp. Since no force sensors are fixed to the claws, we introduced a new variable, called "grasp depth", to represent the grasping effect on the object. The robot grasping strategy diagram is constructed as follows: First, grasp quality is predicted using a linear array network (LAN) and global visual images as inputs. The conditions such as grasp width, depth, position, and angle are also predicted. Second, with the grasp width and depth of the object determined, dynamic movement primitives (DMPs) are employed to mimic human grasp actions with varying velocities. To further enhance grasp quality, a final action adjustment based on tactile detection is performed during the near-grasp time. The proposed strategy was validated through experiments conducted with a Franka robot with a self-designed gripper. The results demonstrate that robot grasping test achieved an increase in the grasping success rate from 82% to 96%, compared to the results obtained by pure LAN and constant grasp depth testing.**

***Index Terms*— Force and Tactile Sensing, learning from demonstration, exoskeleton, data-driven human modelling, robot grasping.**

I. INTRODUCTION

Human skill modelling is a hot research topic in recent years and has been studied in neuromuscular control [1], human dynamic motion optimization [2], motion planning [3], [4] and skill transfer for robot manipulation [5], [6]. Skill modelling relies on accurate measuring sensors, e.g., cameras, force sensors, and displacement sensors. Visual-based detection and recognition is a general way for human-like skill modelling, which can also detect details of interactive objects by using Deep Learning methods. The recent development of deep learning has witnessed visual-based robot manipulation applied in various areas, such as Redmon *et al.* [7] pioneered the use of convolutional neural networks (CNN) to predict parameters of the graspable rectangle in an end-to-end manner. By deepening the network layers [8], [9] and introducing more complex feature fusion modules [10], [11], the performance of grasp detection has been successively improved. In addition, the attention mechanisms [12], [13] were integrated to focus the models on features that are beneficial for grasp detection.

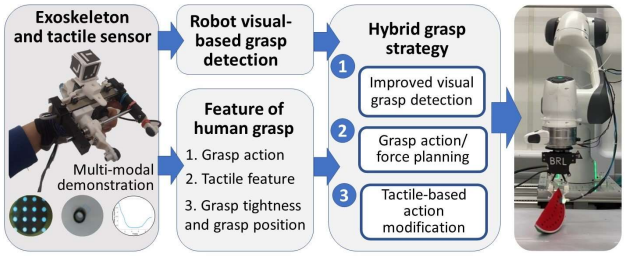


Figure.1 The framework and contributions of the proposed visual-tactile - based robot grasping strategy and skills learning from demonstration

But, as stated in [14], visual-based detecting methods lose detailed information about dense local deformation around the contact areas, leading to the differences between the perceived and actual states. Therefore, visual-tactile fusion-based object perception and robot grasping skill learning have attracted much attention in recent years. Researchers used GelSight [14], BioTac [15], GelSlim [16], and flexible tactile sensor array [17], [18], combining with RGB cameras or RGB-D cameras (e.g., Intel Realsense and Kinect V2), to realize autonomous grasping or improve grasp success.

However, the robot grasping skills are designed manually

Manuscript received: March 10, 2023; Revised: May 2, 2023; Accepted: June 20, 2023.

This paper was recommended for publication by Editor Aniket Bera upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported in part by the H2020 Marie Skłodowska-Curie Actions Individual Fellowship under Grant 101030691 and in part by the National Natural Science Foundation of China (Grant No: 62003200)

* These authors contributed to this paper equally.

[1]Zhenyu Lu, Chenguang Yang, Hengtai Dai and Bofang Zheng are with Bristol Robotics Lab at the University of the West of England, Bristol, BS16 1QY, UK. ([†]corresponding author: Chenguang Yang; e-mail: cyang@ieee.org).

[2]Lu Chen is with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan, China.

[3]Haoran Li and Nathan F. Lepora are with Bristol Robotics Laboratory at the University of Bristol, Bristol, BS8 1TW, UK.

[4]Zhou Zhao is with the School of Computer Science, Central China Normal University. 430079, Wuhan, China.

Digital Object Identifier (DOI): see top of this page.

with little consideration of dynamic modification according to real-time tactile feedback from robots. Hogan *et al.* [18] proposed a model-based regrasp policy and the corresponding data-driven grasp quality metrics. The grasp quality is improved by trials before grasping objects. But, according to the experience of humans, grasping involves pre-grasping, pinching, feeling an object slide on fingertips, and stable grasping. Grasping action and force can be adapted temporally to the visual-tactile feedback. However, as far as we know, there is no research on grasp skill learning from human first perspective demonstrations with robot finger-hand system. The necessity of this problem has been recognized by Ong *et al.* [19]. They proposed a near-contact grasping strategy and claimed that '…all of these grasp planners fail to focus on the interaction between robot fingers and the object, as these grasp planners simply close the robot fingers at the same rate…'. Thus, they utilized kinesthetics teaching to allow robots to learn from human demonstrations. But, humans cannot receive real-time haptic/force feedback to make an appropriate response and modify grasping actions.

This paper aims to develop a novel wearable parallel hand exoskeleton capable of capturing visual-tactile information, and motions to model human grasping skills. The exoskeleton is operated by human demonstrators from a first-person perspective, allowing the operators to experience real-time force feedback through their fingers. The embedded tactile sensors can capture tactile images on the claws during the manipulation. Compared with previous hand exoskeleton designs [20]-[22] and the solutions for learning grasping skills from demonstrations, this exoskeleton was specifically designed for demonstrations, aiming to minimize structural and material differences between the demonstrator (exoskeleton) and robot end effector. The paper presents three core contributions.

1) We designed a new fingertip-sized optical tactile sensor and integrate it into the claws of the exoskeleton, which has been developed based on our previous research [23], [24]. This new tactile exoskeleton can capture multi-modal information to describe the manual grasping process performed by operators accurately.
2) We observed and analysed specific characteristics in human grasping demonstrations, including: a) the varying speed grasping actions, b) shearing, torsional effects, and tactile pressure at the fingertips and c) pre-grasp actions in the grasping process (see Fig. 1(b)).
3) Based on the above design and observations, we proposed a novel framework for learning and applying grasping skills, as shown in Fig. 1. This framework incorporates four steps: a) demonstration b) grasp quality detection, c) grasp action planning and execution, and d) tactile-based action modification to improve grasp quality.

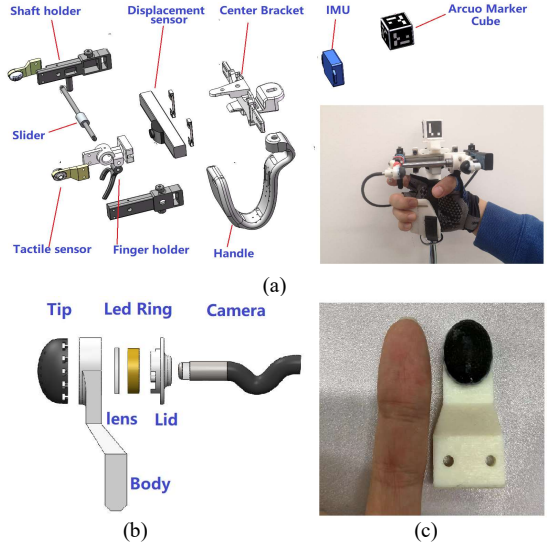


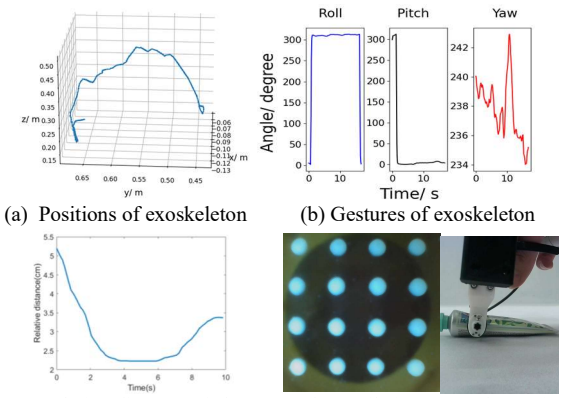
Figure.2 Exoskeleton and designed tactile sensor (a) Exoskeleton structure and wearing effect (b) Structure of TacTip (c) TacTip sensor

II. Tool-like Exoskeleton Design for Generating Demonstration and Data Processing

Different from the demonstrations of human grasping objects using hands directly, this exoskeleton was designed in a shape similar to a parallel robot gripper while the tactile sensors can be fixed to both the exoskeleton and robot gripper, thus the skills learned from human demonstrations can be directly transferred and applied to robots without considerations of material and structural differences between robot gripper and human hand.

Additionally, we considered comfortability and manipulation flexibility in the improvement of our previous works [23], [24], so that the exoskeleton can adapt to operators with different hand sizes and manipulating habits. Before demonstrations, we gave the operators enough adapting time to get used to the exoskeleton till they can smoothly use it.

The tactile sensor was specially designed following [25], [26], which is consisted of a Tip with pins in it (skin and pins are printed with Agilus and markers are printed with Vero white), a fixing body, a lens, a LED ring, a lip and a camera (in Fig. 2(b)). We specially chose a short-sight-distance customized camera (Focusing Range: 5mm, Diameter: 6mm, FOV(D): 90 degrees, 30 FPS, 480p) to develop a finger-size TacTip sensor (in Fig. 2(c)). The camera can detect the distribution of 16 pins embedded in the sensor (in Fig. 3(d)) to detect tactile information. Using the data acquisition software shown in Fig. 5 in [24], we collected multi-modal information to describe demonstrations of human grasping: positions and poses of the hand, the relative distance of gripper claws and the



(a) Positions of exoskeleton (b) Gestures of exoskeleton

(c) Relative distance of claws (d) Tactile images and grasping

Figure 3. Acquired multimodal information of the exoskeleton

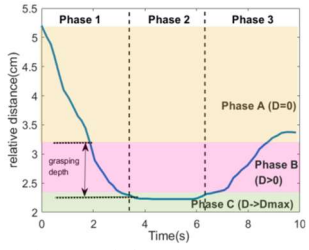
images of optical tactile sensors (in Fig. 3). In our previous research, we have conducted research on robot arm action learning and planning using the hand action and gesture demonstration, but without tactile sensors [23], [24]. In this paper, we focus on in-hand manipulation. Only the movement of the displacement sensor and the tactile images are utilized in this paper.

III. Particularity of Human Grasping Demonstrations

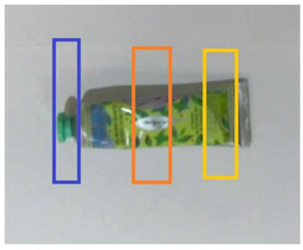
Three volunteers were invited to wear the exoskeleton and demonstrated the grasping of 10 objects, as shown in Fig. 10 (b). Before the demonstration, the volunteers underwent full adaptation to the device and completed basic tests. Initially, a total of 300 global-view images were captured for each object. Then, we labelled the potential grasp positions and guided the volunteers to wear the exoskeleton and manually grasp at the labelled positions. Alternatively, the volunteers can grasp objects randomly before we label the grasping positions. However, through experiments, we found that the positions are hard to accurately mark compared to the current way. To assess the grasp quality and explore the relations between the grasping actions and visual recognitions, we employed a linear array network (LAN) for visual detection and obtained the following three characteristics through the observations.

A. Characteristics I: Varying-speed grasping action and grasp depth in different grasping phases

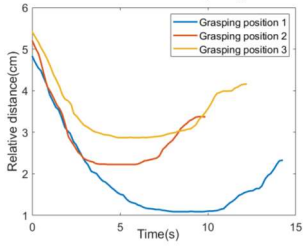
Fig. 4(a) presents the relative distance of the claws of the exoskeleton during a complete human grasp demonstration, which is consisted of three distinct temporal phases over time: Phase 1 represents grasping the object, Phase 2 represents holding the object and Phase 3 represents releasing the object. Furthermore, we introduced a new criterion for grasping detection, named grasp depth D, which takes into account the deformation of the tactile sensor as well as the pressing/grasping direction. This criterion is different from the traditional concept of "grasp width" used in visual-based grasping. The D divides the grasping process into two stages:



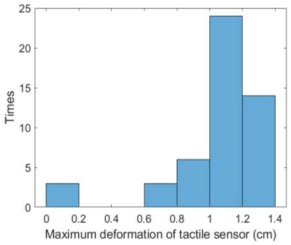
(a) Distance of claws in a grasp case



(b) Object grasp positions

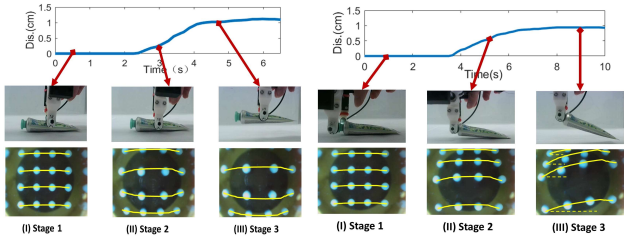


(c) Claws distance of multi-grasp



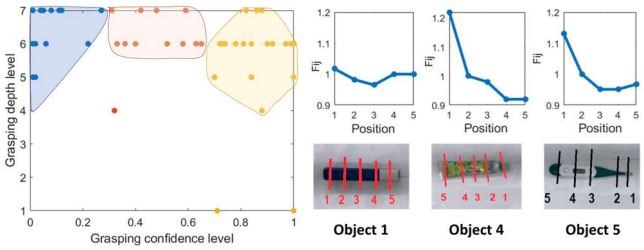
(d) Times of different grasp depths

Figure 4. Grasping actions in human demonstrations



(a) Center grasping position (b) Side grasping position

Figure 5. Images of TacTip with different grasp positions



(a) Grasping level changes with different grasp quality (b) Grasp deformation in different positions on the same object

Figure 6. Grasp depths with different qualities and positions

Phase A, $D = 0$, which indicates no contact with objects and no deformation, and Phase B and C, $D > 0$, representing there is a contact between the object and the tactile sensor. The last stage can be further subdivided according to the grasp depth achieved through either action planning (Phase B) or action modification with feedback from the tactile sensor (Phase C).

It should be noted that the grasp depth D encompasses the combined deformation of both objects and the tactile sensors. Since the tactile sensor is generally much softer than the object, we can assume that D represents the level of deformation in the grasping direction of the tactile sensor after contact with the object. Additionally, Fig. 4(a) reveals that the velocity varies during different phases of the grasping process. In Phase A, the claws approach the contact surfaces at an approximately constant speed. During Phase B, as the object surface come into contact, the grasping speed decreases gradually toward zero along with the increase of contact force. In Phase C, the gripper performs final modifications by increasing the contact force to prevent possible sliding and shearing situations, thus stabilizing the grasped objects, as shown in Characteristics II.

We further analysed the actions in different grasping cases as shown in Fig. 4(b). Their grasping motions are presented in curves in Fig. 4(c). These curves exhibit a similar shape to the curve in Fig. 4(a), except that the grasp width differs. Notably, the thinnest part (indicated by a blue square in Fig. 4(b)) takes more time to be grasped and has a smaller relative distance until the object is fully grasped. This observation reveals that varying-speed grasping is a common phenomenon among humans. Additionally, we investigated the grasp depths for different grasping cases in Fig. 4(d). There were 50 cases used for the analysis, 50% of which had a grasp depth between 1.0 cm and 1.2 cm, which is referred to as a general grasp. The number of looser grasps (grasp depth < 1.0 cm) is less than tighter grasps (grasp depth > 1.2 cm). Additionally, two failures occurred during the demonstration (grasp depth < 0.2 cm).

B. Characteristics II: Shear, twist and press in different grasp phases

Fig. 5 displays the images captured by the camera inside the tactile sensor during different stages of the object grasping at different positions. In Fig. 5(a), it is evident that the object was stably grasped at the centre of the object, with a grasp depth of 1.1cm approximately. During the grasping process, there was no rotation observed, so the distribution of markers was expanded throughout the grasping process consistently until the object was fully grasped. Fig. 5(b) shows an instance of edge grasping. In this case, due to the instability, a noticeable twist occurs on the object at Stage 3, as reflected by the changes in row spacing and angular markers in the haptic images, indicating shearing and torsion effects. The posture of the object differs between the stages of partial grasp (Stage 1 and Stage 2) and the fully grasped state (Stage 3). This distinction underscores the need for a final tactile detection and action modification in Phase 3, as depicted in Fig. 4(a).

C. Characteristics III: Relationship between grasping position and grasp depth.

Humans can approximately plan their actions and anticipate the expected grasping effect and force upon seeing an object. While the traditional visual-based grasping effect depends on the force sensor measurement. This characteristic is to explore the relationship between grasping action planning and visual detections. First, we made a quantization of the grasp depths for the investigation in Fig. 4(d) and divided 7 levels in Fig. 6(a) with an interval of 0.2 cm. The X-axis represents the grasp quality predicted by the LAN method in Section IV. The scatter points are classified into 3 classes based on different quality levels:1) [0, 0.3), 2) [0.3, 0.7], and 3) (0.7, 1] and show with different colours. The contours of the classes are outlined as well. The blue nodes with an inverted triangle contour, representing the predicted low-quality grasp positions require more higher-level (Level 7) grasps. The contours of the middle and high-quality positions (red and orange nodes) take the shape of a rectangle and a polygon, respectively, indicating they require relatively lighter grasping force/depth. However, this finding is derived from the limited demonstration samples and cannot be used as a definite principle for robot grasping.

Additionally, the grasping force and depth can be partly determined by the weight of the object. Therefore, a new variable F_j^i is introduced to assess the relative grasp effect on the kth grasping point of the jth object, which has the highest predicted grasp quality based on the deep learning method:

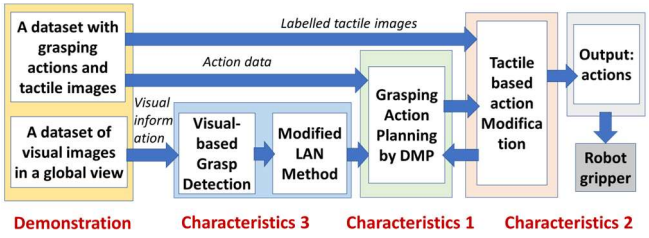


Figure 7. Diagram of the proposed grasping strategy

$$F_j^i = D_j^i / D_j^k,\ i,j \in N^*: 1 \le i \le M, i \ne k \tag{1}$$

where D_j^i represents the grasp depth of the tactile sensor at the ith position. Fig. 6(b) illustrates the values of $F_j^i, 1 \le i \le 5$ for three objects. Object 1 has a value of $F_1^i \in [0.96, 1.02]$, and its varying range is much smaller compared to those of Object 4 ($F_4^i \in [0.92, 1.2]$) and Object 5 ($F_5^i \in [0.95, 1.14]$), This is because Object 1 is a lightweight box with an average width. Comparatively, the grasping points located on the edges and bounding areas of the objects are expected to require a larger force to maintain stability and prevent slipping and shearing on the object's surface.

IV. Visual-Tactile Robot Grasping Based on Human Grasping Skill Modelling

A. Overview

The new grasping strategy is proposed in Fig. 7 based on the analysis of characteristics in Section III. The demonstrations were collected and preliminarily processed to get two datasets: one is for visual-based detection and the other consists of grasping actions and tactile images. We improved LAN, a deep-learning method for visual recognition, to predict the variables including position, angle, width and depth, in grasping actions planning based on Characteristics 3 concluded in Section III. The variables start the action planning process by the grasping skills learned by Dynamic Movement Primitives (DMP) based on the human demonstration and Characteristics 1. Finally, the grasping actions at the near-grasp time are modified to ensure grasping stability based on the tactile feedback in real time (Characteristics 2). The modules of the strategy in Fig. 7 are presented separately in the following subsections.

B. Visual-based Grasp Detection and Its Improved Model

To evaluate the grasp qualities of different positions on the object, we used the LAN from our previous work for low-light image enhancement [27] to achieve pixel-wise grasp detection. This network proposes a linear array network to construct 3D weights with 2D feature encodings and combines global and local information to learn specific grasping features. The framework of the LAN-based grasp detection model is shown in Fig. 8. For the input RGB image I, it successively flows through the shallow feature extraction module F_{sfe}, which is composed of a convolutional layer, a linear array module F_{lar}

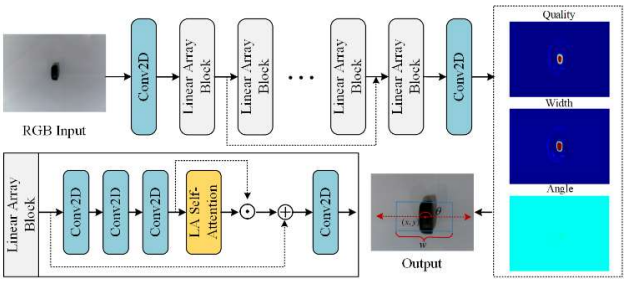


Figure 8. Framework structure of LAN-based grasp detection model. Please refer to [27] for details.

(seven linear array blocks) and a feature mapping module F_{fma} (a convolutional layer), to generate three grasping feature maps with respect to quality, width and angle.

The method was improved by adding a new criterion about grasp depth, where the input images are labelled by the factor F_j^i in (1) to present the impact of the non-averageable width of the object on the grasp depth. It is obvious that F_j^i is a non-dimensional factor corresponding to Characteristics 3. It has two functions, one of which is to correct the grasp quality map. Through human demonstrations, we observed that even for positions with low confidence, the gripper can grasp the object up by exerting a larger force. Therefore, if Q is set to be a grasp quality map learned by LAN, we can correct Q_j^i, the ith test quality map of the jth object using F_j^i, to achieve $\bar{Q}_j^i$:

$$\bar{Q}_j^i = Q_j^i \times F_j^i, \quad if \quad Q_j^i < \hbar \tag{2}$$

where $\hbar$ is a threshold value for judging if the grasp quality should be improved by modifying the grasp depth. After searching the local peaks of the improved quality map $\bar{Q}$, the grasp position with the highest confidence level is regarded as the predicted grasp.

$$g = f^{\bar{Q}}(F_{fma}(F_{lar}(F_{sfe}(I)))) \tag{3}$$

where g is the predicted grasp composed of six parameters $\{x, y, w, \theta, q, f\}$ representing the spatial positions, width, angle, quality and a factor with respect to the items in F_j^i in (2). $f^{\bar{Q}}$ is a function of searching local peaks on $\bar{Q}$. The smooth L1 loss is used to measure the difference between predictions and the ground truth grasps. The other function of F_j^i is to make grasping action planning. A generally desired grasp depth $\bar{D}_j$ is set first (e.g., 1.2 cm in Fig. 4(d)), which takes up the majority of grasping cases. The potential grasping positions, $\bar{D}_j^i$ are predicted by $D_j^i = \bar{D}_j \times F_j^i$.

C. Grasp Action Modelling and Planning using DMP

Following Characteristics I, we can observe that human grasping velocity varies throughout the phases from pre-grasping to the final releasing. Firstly, the claws approach the contact surface at an almost constant velocity. Upon contact with the object, the grasping velocity gradually decreases along with the larger deformation of the tactile sensor and increasing contact forces. To achieve human-like grasping actions, dynamical movement primitives (DMPs) were used, which suit both one demonstration and multiple demonstrations for robot grasping action learning and generalization. The action starts from the initial grasping pose and ends with the position with the largest grasp depth.

With the knowledge of the start x_0 and the end g, the DMP function in (4) is learned from demonstrations as:

$$\begin{cases} \tau \dot{v} = \alpha_z \left(\beta_z \left(g - x \right) - v \right) + f\left(s \right) \\ \tau \dot{x} = v \end{cases}, \tag{4}$$

where x and $\dot{x}$ are position and velocity, x_0 is the start of x, and τ with a scaling factor. $f\left(s \right) = \mathbf{W}^T \mathbf{\Psi}(s)$ is a nonlinear term, where $\mathbf{W} = \left[w_1, w_2, ..., w_n \right]^T$, $\mathbf{\Psi}(s) = \left[\psi_1, \psi_2, ..., \psi_n \right]^T$, and

$$\psi_i = \frac{\exp(-h_i (s - c_i)^2) s}{\sum_{i=1}^{n} \exp(-h_i (s - c_i)^2)}, \tag{5}$$

where c_i and $h_i > 0$ are the centres and widths of the Radial Basic Functions, and $\alpha_z, \beta_z > 0$ are coefficients. s is a phase variable to ensure the dependency of $f\left(s \right)$ out of time, which is expressed by a canonical system $\tau \dot{s} = -\gamma s, \gamma > 0$. Setting the target values of $f\left(s \right)$ calculated the ith demonstration x_i as:

$$\begin{cases} f_i^{Tar} = \left(\tau \dot{v}_i - K\left(g_i - x_i \right) - D v_i \right) \\ \tau \dot{x}_i = v_i \end{cases}, \tag{6}$$

then $f\left(s \right)$ is achieved by multiple demonstrations ($N > 1$) or a single demonstration ($N = 1$), by minimizing

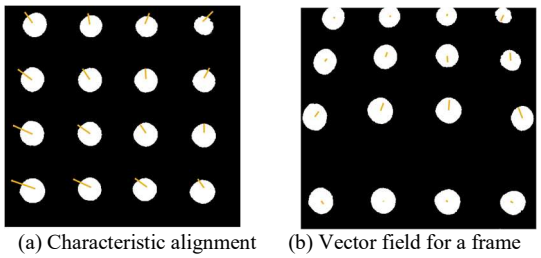
$$\min \left(\sum_{k=1}^{N} \left(f_i^{Tar} - f\left(s \right) \right)^2 \right) \tag{7}$$

to achieve the parameter vector $\mathbf{W}$. The learned trajectory x in (4) is then generalized for the new case with different x_0, g and τ in [28], [29]. Since continuous speed modification cannot be realized in servo motor control. In the experiment, we simplified the generalized trajectory into two line segments between the start and the target.

D. Tactile-based Grasp Modification

Tactile-based action adjustment is the last step. The images acquired by the optical tactile sensor are binarized as shown in Fig. 9. The K-means clustering is then used to get the central position of each pin between consecutive frames. After calculating the vector field of the pins' movement, the slipping and shearing actions can be detected based on our previous research [30]. In this work, considering there is a limited number of pins, we proposed a simple metric to judge if the gripper should close more to stabilise the grasping state. Setting the movement of pins is expressed by $\mathbf{V}_j = [x_j, y_j]^T$ and $\theta_j = atan(y_j / x_j), j = 1, 2, ..., 16$. The tactile detection is performed for every 5 frames. If any condition in (8) is satisfied:

$$\left| \sum_{j=1}^{16} x_j \right| > \Delta x \,, \left| \sum_{j=1}^{16} y_j \right| > \Delta y \,, \left| \sum_{j=1}^{16} \theta_j \right| > \Delta \theta \tag{8}$$



(a) Characteristic alignment of pins when slip occurs (b) Vector field for a frame where the object is held

Figure 9. Detection of marker motions of tactile sensors

A sliding or shearing action would be reckoned to occur on the contact surface of the object, where Δx , Δy and $\Delta\theta$ represent the thresholds. Then the closing distance of claws is increased with δ to generate a larger force for every detection period. Since the tactip currently only contains 16 markers, for future research, we plan to explore image-based deep learning methods for a new tactile sensor that incorporates more markers with increased density.

V. Experiments

A. Data Collection and Registration

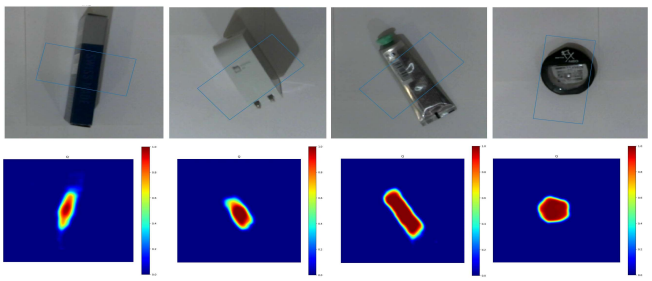
Data collection was conducted using the platform depicted in Fig. 10(a), which includes an exoskeleton, a global camera (3840 x 2160 pixels, 30fps, Autofocus 100-degree lens), a side web camera (1080p, 30fps, ASHU) for judging object pickup, a laptop for data acquisition, and a frame for manipulation. The objects used for grasping demonstrations are shown in Fig. 10(b). We compiled a dataset consisting of 300 labelled images captured from the global view for visual detection, a dataset of grasping actions (e.g., hand motions, gestures, and movements of the claws, but only the last is used in this experiment), and videos recorded from the tactile sensors and the side camera. The frames were sampled from the videos at the same rate as the action sampling and aligned with the action dataset. The duration of object grasping and placing was determined and labelled manually by human annotators.

(a) Demonstrating platform (b) Grasping objects

Figure 10. Demonstrating platform and grasping objects

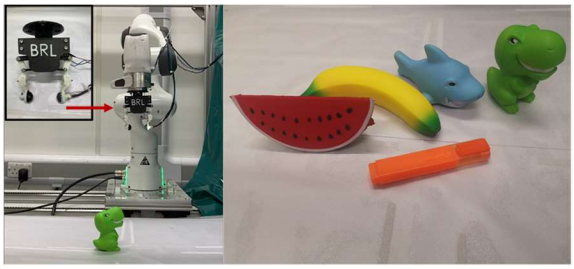
B. Visual-based Grasp Detection Results

Considering the low capacity of our visual/tactile dataset, we initially pre-trained the visual grasp detection model using the Cornell Grasping Dataset [31] and fine-tuned the model parameters using our collected images. All experiments were conducted on a single NVIDIA RTX 3090 GPU with the operating system of Ubuntu 20.04. In order to evaluate the

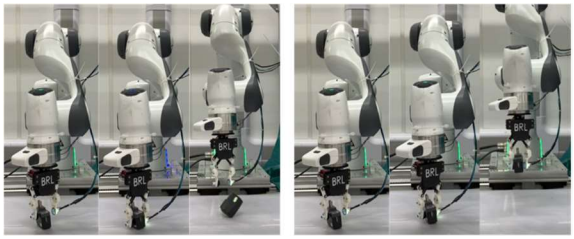


(a) image 1 (b) image 2 (c) image 3 (d) image 4

Figure 11 Qualitative results of grasp detection model. First row: predicted grasps, second row: grasp quality maps.



(a) Grasping platform (b) New added grasping objects for test



(c) Failed grasp using LAN (d) Grasp using the proposed method

Figure 12. Platform and experiments of robot grasping objects

detection accuracy, we adopted the commonly used rectangle metric [32]. A predicted grasp was considered correct if it satisfied the following two criteria:

1) Jaccard index J_{ind} is defined as the intersection between the predicted grasp g_p and grounds truth grasp g_{gt} divided by their union, which is larger than 0.25:

$$J_{ind} = \frac{|g_p \cap g_{gt}|}{|g_p \cup g_{gt}|}, J_{dif} = |\theta_p - \theta_{gt}| \quad (9)$$

2) The difference J_{dif} between θ_p and θ_{gt} is less than 30^o.

During the training process, we randomly divided 90% of the total images as training samples and used the remaining 10% for testing purposes. Data augmentation techniques such as image zooming, and rotation were employed. The initial learning rate was set to 0.001, the batch size was set to 8, and the Adam algorithm was utilized for optimizing the model parameters. The LAN model achieved a final grasp detection accuracy of 87.0% on our collected dataset, while the improved model further increased this value to 90% by multiplying the factor F_j^i in (2). Fig. 11 showcases some qualitative detection results, demonstrating the robustness of the grasp detection model to variations in object scale and spatial position. Moreover, it can be found that the detected grasps concentrate around the geometric centres of the objects predominantly. The grasp quality in some edging areas is improved to some extent, which facilitates the grasping in real-world scenarios by combining pre-grasp actions with tactile-based action modification.

C. Grasping Verification based on Robot Platform

The grasping action planning and tactile-based modification are performed according to the steps outlined in Fig. 7 with the parameters: $\alpha_z = 100, \beta_z = 25, \gamma = 0.9$ and $\tau = 0.1$. Considering the gripper is driven by a servo motor and gears, we further approximated the generalized gripper movement by two line segmentations. Set $\Delta x = 50$, $\Delta y = 50$, $\Delta\theta = 180$ and $\delta = 1mm$ in Section IV. D, the grasping verification was performed based

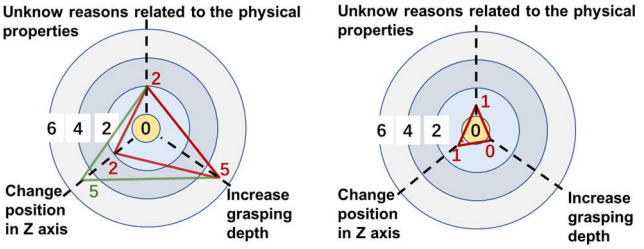
on the platform with a Franka robot, a global camera, and a self-designed gripper (Fig. 12(a)). The added objects are shown in Fig. 12(b). Since we only acquired RGB-global view images, LAN-based robot grasping acted with a constant grasp depth. The comparative grasping results are presented in Fig. 12 and Table I, where for each method, the old and newly added objects are mixed and randomly grasped 50 times. We can see from Table I that using the proposed strategy, the grasp successful rate taken on real robot system increases from 82% to 96% through pre-grasping planning and modifications, which proves that the proposed grasp strategy can improve grasping stability and avoid potential failures to some extent.

TABLE I. GRASP ACCURACY ON REAL OBJECTS

Method	Grasp (Successful Times/Times)
LAN [27]	41/50
Our method	48/50

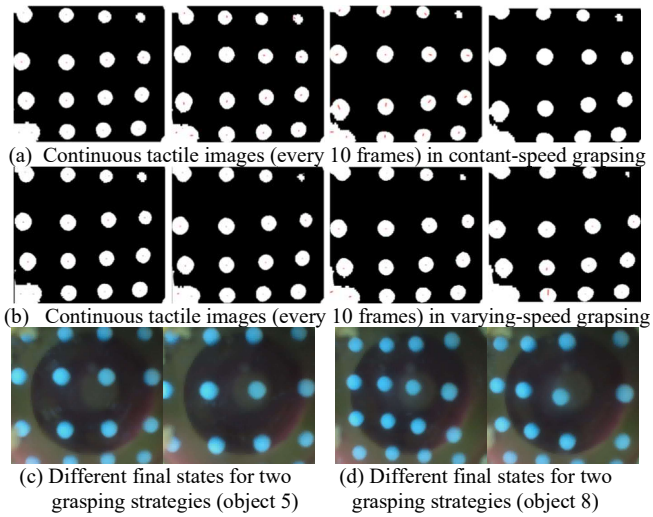
D. Failure Reasoning and Discussion

Although the success rate of grasping has increased, there are still instances of failure due to various reasons. Fig. 12(c) and Fig. 12(d) display the failed and successful grasp cases of a headphone box separately. Even under the same conditions, it is possible to encounter failures during the grasping process. We conducted a comprehensive analysis based on the observed failure cases in both the LAN method and our proposed method. We propose the following three criteria for evaluation: 1) if the object can be grasped by modifying the grasping positions in the Z-axis through the robot's movement, 2) if the object can be grasped by increasing grasp depth, and 3) the unknown reasons related to the physical properties, such as the surface stiffness or the shapes of the object which cannot be addressed by 2D grasping strategies. From Fig. 13, we can see that the grasping failures in the cases performed based on LAN can be mostly corrected by increasing the grasp depth (5 cases) and changing the grasp positions in the Z-axis (5 cases) through regrasping. There are also 3 cases that could be solved both ways, as indicated by the difference between the green lines and the red lines shown in Fig. 13. However, there are 2 cases that cannot be addressed by any actions using the robot and gripper system. For our proposed method, there is 1 failed case that can be solved by changing the grasp position in the Z-axis, and 1 case that cannot be solved at all. As the failed cases are determined based on visual characteristics, the future research will add a weighting platform to use weight changes for judging if the objects are grasped successfully.



(a) Failure reasoning of LAN-based grasping (b) Failure reasoning of grasping based on the proposed method

Figure 13. Failure reasoning of results in comparative methods



(a) Continuous tactile images (every 10 frames) in contant-speed grapsing

(b) Continuous tactile images (every 10 frames) in varying-speed grapsing

(c) Different final states for two grasping strategies (object 5) (d) Different final states for two grasping strategies (object 8)

Figure 14. Tactile images of constant and varying speed grasping

Another question pertains to the improvement of grasping performance through grasping actions learning from human demonstrations. In this work, human demonstrations serve two purposes. Firstly, they are used to predict the grasp depth by examining the correlations between human actions and visual detection results. Human demonstrations contribute a new evaluation principle, grasp depth, for labelling visual images. Secondly, they enable learning from action data to acquire a grasping skill that resembles human performance, accounting for variations in velocity. Here, we compared the effects of using a constant grasping speed to approach the deepest grasp depth with the results obtained from utilizing varying speeds. Fig. 14(a) shows the tactile images captured every 10 frames in a constant speed grasping and Fig.14 (b) shows those in a varying speed grasping. We can see that the displacements of the centre positions of the markers in Fig.14 (b), indicated by the red line segments, are much shorter in the adjacent pictures. Thus, the varying speed grasping presents the contact process more clearly and provides enough time for tactile recognition.

Additionally, we performed these different grasping strategies on two other grasp tasks separately and find that the final states of the markers are different. The deformation degree of the tactile sensor is relatively more serious in varying-speed grasping, as shown in the right figures of Fig. 14(c) and Fig. 14(d). However, the grasping stability for the tests shows no obvious difference, as the objects are held tightly. From our perspective, varying speed grasping is more suitable for manipulating fragile objects with slow, soft, and sufficient contact, while also balancing grasping efficiency through the initial fast-approaching phase.

VI. CONCLUSIONS AND FUTURE WORK

This paper presents a novel exoskeleton design and a visual-tactile grasping model inspired by the human grasping process. To this end, the new parallel hand exoskeleton equipped with optical tactile sensors can measure multimodal information and provides timely force feedback to human fingers during a grasping task. After analyzing the demonstrated data, we concluded three characteristics of human grasp and use them to build a human grasping skill model and related robot grasping strategy by combining imitation learning (DMP), deep learning (LAN) and tactile-based motion planning. The experiments were conducted on a Franka robot to prove that the proposed strategy can significantly improve grasp quality.

Compared with visual-based grasping, we fully employed the perfect dynamic detection performance of TacTip sensors

for grasping based on human demonstrations, which is an integration of offline and real-time robot motion planning to compensate for the errors of visual prediction and autonomous robot grasping. Future work will focus on these two directions:

1) Functional grasp. Like using a hammer or a key, the high-quality grasping points may be not the best positions for using objects. In our previous research [33], we studied robot skill learning and transfer for multi-tool use cases. How to improve the grasp affordance, and then grasp tools and use the tools to complete a task is the following work after object grasping. Creating a new grasp detection metric from the perspective of tool-use is essential. Undoubtedly, learning from demonstration (LfD) is an optional way and our developed exoskeleton is a useful tool to acquire grasping and manipulating information.

2) Hybrid visual-tactile grasp detection metric and an online visual-tactile-based grasp motion control. In addition to the research presented in this paper, we created an extra database of tactile images categorized by quantized grasp depth, similar to the grasping level division outlined in Section III. This enables us to use deep learning to extract tactile features as feedback for real-time grasp motion control. Furthermore, the process of adaptively fusing data from different modalities requires further investigation.

REFERENCES

[1] A. Seth *et al.,* "OpenSim: Simulating musculoskeletal dynamics and neuromuscular control to study human and animal movement*,* " *PLoS Comput. Biol.*, vol. 14, no. 7, 2018.

[2] G. Schultz and K. Mombaur, "Modelling and optimal control of human-like running," *IEEE/ASME Trans. Mech.*, vol. 15, no. 5, pp. 783-792, 2010.

[3] E. Yoshida, I. Belousov, C. Esteves and J.P. Laumond, "Humanoid motion planning for dynamic tasks," *5th IEEE-RAS Int. Conf. Humanoids.*, 2005, pp. 1-6.

[4] C. Fang, X. Ding, C. Zhou, and N. Tsagarakis, "A2ML: A general human-inspired motion language for anthropomorphic arms based on movement primitives," *Robot. Auton. Syst*., 111, pp.145 -161, 2019.

[5] C. Fang, G. Rigano, N. Kashiri, A. Ajoudani, J. Lee and N. Tsagarakis, "Online joint stiffness transfer from human arm to anthropomorphic arm," *2018 IEEE Int. Conf. Syst., Man, Cybern.*, 2018, pp. 1457-1464.

[6] L. Peternel, T. Petrič, E. Oztop, and J. Babič, "Teaching robots to cooperate with humans in dynamic manipulation tasks based on multi-modal human-in-the-loop approach," *Auton. Robot.,* vol. 36, pp. 123 -136, 2014.

[7] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," *2015 IEEE Int. Conf. Robot. Automat.*, 2015, pp. 1316-1322.

[8] F. -J. Chu, R. Xu and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3355-3362, 2018.

[9] L. Chen, P. Huang, Y. Li and Z. Meng, "Edge-dependent efficient grasp rectangle search in robotic grasp detection," *IEEE/ASME Trans. Mechatron.*, vol. 26, no. 6, pp. 2922-2931, Dec. 2021.

[10] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," *2018 IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2018, pp. 7223-7230.

[11] D. Park, Y. Seo and S. Y. Chun, "Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module," *2020 IEEE Int. Conf. Robot. Automat.*, 2020, pp. 9397-9403.

[12] S. Kumra, S. Joshi and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," *2020 IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2020, pp. 9626-9633.

[13] S. Yu, D. -H. Zhai, Y. Xia, H. Wu and J. Liao, "SE-ResUNet: A novel robotic grasp detection method," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5238-5245, 2022.

[14] W. Yuan, S. Dong, and E. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[15] K. Ganguly, B. Sadrfaridpour, P. Mantripragada, N.J. Sanket, C. Fermüller, and Y. Aloimonos, "Grasping in the dark: zero-shot object grasping using tactile feedback," arXiv:2011.00712 [cs], 2020.

[16] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson and A. Rodriguez, "GelSlim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," *IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2018, pp. 1927-1934.

[17] T. Li et al., "Robot grasping system and grasp stability prediction based on flexible tactile sensor array," *Machines*, vol. 9, no. 6, p. 119, Jun. 2021.

[18] F. R. Hogan, M. Bauza, O. Canal, E. Donlon and A. Rodriguez, "Tactile regrasp: Grasp adjustments via simulated tactile transformations," *2018 IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2018, pp. 2963-2970.

[19] Y. H. Ong, J. Morrow, Y. Qiu, K. Gupta, R. Balasubramanian and C. Grimm, "Near-contact grasping strategies from awkward poses: When simply closing your fingers is not enough," *2019 IEEE/RSJ Int. Conf. Intell. Robots and Syst.*, 2019, pp. 646-651.

[20] T. du Plessis, K. Djouani, and C. Oosthuizen, "A Review of Active Hand Exoskeletons for Rehabilitation and Assistance," *Robotics*, vol. 10, no. 1, p. 40, 2021.

[21] M. Sarac, M. Solazzi and A. Frisoli, "Design requirements of generic hand exoskeletons and survey of hand exoskeletons for rehabilitation, assistive, or haptic use," *IEEE Trans. Haptics.*, vol. 12, no. 4, pp. 400-413, 2019.

[22] M. Cempini, M. Cortese and N. Vitiello, "A powered finger–thumb wearable hand exoskeleton with self-aligning joint axes," *IEEE/ ASME Trans. Mechatron.*, vol. 20, no. 2, pp. 705-716, 2015.

[23] H. Dai, Z. Lu, M. He and C. Yang, "Novel gripper-like exoskeleton design for robotic grasping based on learning from demonstration," *2022 27th Int. Conf. Autom., Comput.*, 2022, pp. 1-6.

[24] H. Dai, Z. Lu, M. He, and C. Yang, A gripper-like exoskeleton design for robot grasping demonstration. *Actuators*, vol. 12, no. 1, p. 39, 2023.

[25] N. F. Lepora, "Soft biomimetic optical tactile sensing with the tacTip: A review," *IEEE Sens. J.*, vol. 21, no. 19, pp. 21131-21143, 2021.

[26] B. Ward-Cherrier *et al.*, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft Robot.*, vol. 5, no. 2, pp. 216-227, 2018.

[27] K. Wang, Z. Cui, J. Jia, et al. "Linear array network for low-light image enhancement," arXiv:2201.08996 [cs], 2022.

[28] Z. Lu, N. Wang, M. Li and C. Yang, "Incremental motor skill learning and generalization from human dynamic reactions based on dynamic movement primitives and fuzzy logic system," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 6, pp. 1506-1515, 2022.

[29] Z. Lu, N. Wang and C. Yang, "A constrained DMPs framework for robot skills learning and generalization from human demonstrations," *IEEE/ASME Trans. Mech..*, vol. 26, no. 6, pp. 3265-3275, 2021.

[30] J. W. James, N. Pestell and N. F. Lepora, "Slip detection with a biomimetic tactile sensor," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3340-3346, 2018.

[31] I. Lenz, H. Lee and A. Saxena, "Deep learning for detecting robotic grasps", *Int. J. Robot. Res.*, vol. 34, no. 4-5, pp. 705-724, 2015.

[32] Y. Jiang, S. Moseson and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," *2011 IEEE Int. Conf. Robot. Automat.*, 2011, pp. 3304-3311.

[33] Z. Lu, N. Wang, M. Li and C. Yang, "A novel dynamic movement primitives-based skill learning and transfer framework for multi-tool use," *2022 IEEE 17th Int. Conf. Control. & Autom.*, 2022, pp. 1-8.