ROBOT GRASPING AND MANIPULATION COMBINING VISION AND TOUCH

Zihao Ding,* Guodong Chen,* Zhenhua Wang,* and Lining Sun*

Abstract

Humans tend to instinctively integrate information from various senses such as vision and touch to accomplish dynamic adjustments when grasping and manipulating objects. The current robot manipulation tasks largely depend on visual guidance, resulting in the inability to cope with the contact activities demanding precise control. Vision and touch are separate processes in recent fusion methods, which are far from what happens in the human brain. The fusion is also unable to solve the problem of damage to the objects during initial contact. Therefore, a pre-grasp network based on the fusion of visual detection and tactile prior knowledge is proposed in this paper, which combines visual image and tactile experience to reach fast pre-grasp for robots. Then, using the optimal self-search of the time step, a tactile network is built to automatically adjust the time step and output particular grasp hardness and grasping state for the object. Finally, the dexterous robot hand can be constantly controlled for steady grasping utilising the force/position control algorithm. Experiments show that this method is useful for the robot to complete the stable grasp and manipulation of different objects.

Key Words

Robot grasping, neural network, tactile prior knowledge, visual-tactile fusion, optimal self-search, force/position control

1. Introduction

As an important tool to improve the intelligent and manipulative levels of the robot, the multi-sensory dexterous robot hand has become one of the most promising researches in the robot field [1]. Predicting the grasping results of dexterous hands [2] is crucial to realise robot manipulation [3], which can assist robots in developing grasping strategies. Many studies have proposed various methods to accomplish this task [4], [5]. Early studies focused on physical modeling of the grasping objects, grippers, and environment, typically using visual [6], [7] or depth observations [8]. The inclusion of tactile input to this task, according to recent studies [9], [10], might enhance prediction accuracy significantly. Li and Adelson [11] demonstrated a vision-based tactile device with GelSight material as a transparent elastomer in 2012. Hai *et al.* [12] constitutes a tactile perception system by fusing the electromyography signal controller of the hand, the vibrotactile system, the torque sensor, and the motor driver. The vibrotactile system can provide a sense of grasping force to the subject. Using BioTacs, Chebotar *et al.* [13] presented a framework for relearning grasping behavior based on tactile data. Several works [14]–[17] suggested the use of tactile sensors to assess grip stability. Veiga *et al.* [18] extracted features from tactile signals to detect/predict slip to adaptively adjust the grasp force. Han *et al.* [19] designed a multi-modal CNN model, which could obtain the hardness, thermal conductivity, roughness, and texture features to modify the grasping strategy. Li *et al.* [20] focused on finding the appropriate pressure distribution to realise the stable grasping operation of a variety of objects. James and Lepora [21] demonstrated the slip detection capabilities of robotic hands by using support vector machines.

However, we should not entirely isolate vision and touch. Humans frequently integrate touch, vision, and even hearing in manipulation tasks [22]. The study has shown that the human brain utilises a multisensory model [23]. Employing model-based techniques, researchers have developed robotic systems that combine visual and tactile information for grasping. Wallhoff *et al.* [24] introduced a teachable hybrid assembly system that was capable to process voice, gaze, and tactile interaction channels. Wang *et al.* [25] proposed a novel paradigm that efficiently perceives accurate 3D object shapes by incorporating visual and tactile observations, as well as priorknowledge of common object shapes learned from large-scale shape repositories. Guo *et al.* [26] generated the initial grasp rectangle by visual perception, and then evaluated the grasp quality by tactile information and strain gauge. Cui *et al.* [27] proposed a visual-tactile fusion learning method based on the self-attention mechanism. Calandra *et al.* [10] predicted grasp success probability based on tactile readings, RGB image, and a regrasping action for a two-fingered gripper. While the aforementioned method of

* School of Mechanical and Electric Engineering, Jiangsu Provincial Key Laboratory of Advanced Robotics, Soochow University, Suzhou, China; e-mail: zhding@stu.suda.edu.cn; guodongxyz@163.com; wangzhenhua@suda.edu.cn; lnsun@hit.edu.cn
Corresponding author: Guodong Chen

Recommended by Hicham Chaoui
(DOI: 10.2316/J.2024.206-0964)

using both visual and tactile information simultaneously can enhance perceptual abilities during a task, it fails to take into account the inherent relationship between visual and tactile information.

Caporali *et al.* [28] presented the fusion method between the shape estimation provided by the vision system and the one provided by the tactile sensor to grasp the cable. They investigated the complementary relationship between haptics and vision under occlusion conditions. However, the tactile sensors used were specific to the grasping task of the cable and could not be applied to other robotic tasks. Han *et al.* [29] proposed a transformer-based robotic grasping framework for rigid grippers that leverage tactile and visual information for safe object grasping. Kanitkar *et al.* [30] used the tactile and vision data obtained during grasping and moving the object to the holding pose to predict whether the object is stable. Matak and Hermans [31] proposed an approach to grasp planning. The key to the method's success is the use of visual surface estimation for initial planning. The robot then executes this plan using a tactile-feedback controller that enables the robot to adapt to online estimates of the object's surface to correct for errors in the initial plan.

The visual grasping method is difficult to achieve dynamic adjustment. The above grasping strategy of fusing visual and tactile information depends on multiple trials and errors, and the fusion methods are based on their special hardware, which is not scalable. Furthermore, no work has shown the ability to plan, adapt, and ensure precision grasps for multi-fingered hands through joint visual and tactile sensing in the real world [31]. On the one hand, the initial force of the grasp cannot be set leading to the damage of the object during contact. On the one hand, the fusion approach is employed for object recognition and slip detection, not for changing the robot's grasping state.

Therefore, we propose a new visual and tactile fusion solution that overcomes the limitations of traditional methods of applying hardware platforms and solves the initial grasping problem before grasping and the dynamic adjustment problem during grasping. Our primary contributions are three-fold:

- A pre-grasp convolutional neuronal network based on the fusion of visual detection and tactile prior knowledge is proposed to accomplish fast and safe pre-grasp for robots and solve the problem of loss due to missing tactile information in the initial grasping phase. And a new visual-tactile data set and fusion model are established.
- A tactile long short-term memory (LSTM) network based on optimal self-search of the time step (OSLSTM) is built to assess the grasping state and the grasp hardness for the object. And the force/position control algorithm is used to adjust the grasping state in real time based on the attributes of the object.
- The results of the experiments show that this approach can avoid damage and solve the problem of unstable grasping.

The rest parts of this paper are organised as follows. Section 2 introduces the visual and tactile samples in the visual-haptic fusion, and expounds the structure of the grasping network. Then, Section 3 describes the grasp state evaluation model and adjustment mode. The experimental results are shown in Sections 4, and 5 makes a conclusion for this paper.

2. A Pre-Grasp Network based on the Fusion of Visual Detection and Tactile Prior Knowledge

In the task of robot grasping, not only the location information of the target is required, but also the grasping position can be detected. Besides, different forces, positions, and other strategies should be formulated according to the different attributes of the object materials. In current tactile-based grasping methods, the robot continuously attempts and adjusts the force after contacting objects with their gripper, which is time-consuming. It is also easy to harm the target of the initial touch if the initial force was not appropriately adjusted for soft material targets.

Thus, this article develops a pre-grasp network based on the fusion of visual detection and tactile prior knowledge for the three-fingered dexterous hand, as shown in Fig. 1. Firstly, a significant number of grasping experiments is conducted on various objects, yielding visual and tactile data sets. The grasp position and grasp force necessary for a successful grasp task are determined using the grasping outcomes. We build the connection between visual images and grasp attributes using tactile information and the visual network. By inputting the visual image into the trained neural network, we can get grasp information, object categories, and pre-grasp force. In this way, only visual detection is required to successfully grasp the object in the actual test.

2.1 An Optimised Data Set for Visual and Tactile Fusion

2.1.1 Vision: Sample Labels based on Objects Material

We use the Cornell Grasp Dataset as the visual training set, which are comprised of 885 images of 240 different objects. Each image has multiple grasp rectangles marked as success (positive) or failure (negative). There is a close correlation between the material of an object and its grasping method, which has been ignored in previous studies. The dataset comprises samples from six categories, including the most typical shapes of cuboid, cylinder, and sphere, each with two unique properties of soft or hard, resulting in a total of 48 different sample types. The diverse shapes and materials necessitated different grasping techniques, and this categorisation helped to better associate the tactile data with the corresponding shape and material. The examples of dataset classification are presented in Fig. 2.

The grasping labels are represented by six-dimensional data $Q(t,s) = \lambda_t \sum_n^{i=1} \ln(t_i+1) + \lambda_s \sum_n^{i=1} \ln(s_i+1)$, with $Q(t,s) = \lambda_t \sum_n^{i=1} \ln(t_i+1) + \lambda_s \sum_n^{i=1} \ln(s_i+1)$ corresponding to the center coordinate of the grasp rectangle, $Q(t,s) = \lambda_t \sum_n^{i=1} \ln(t_i+1) + \lambda_s \sum_n^{i=1} \ln(s_i+1)$

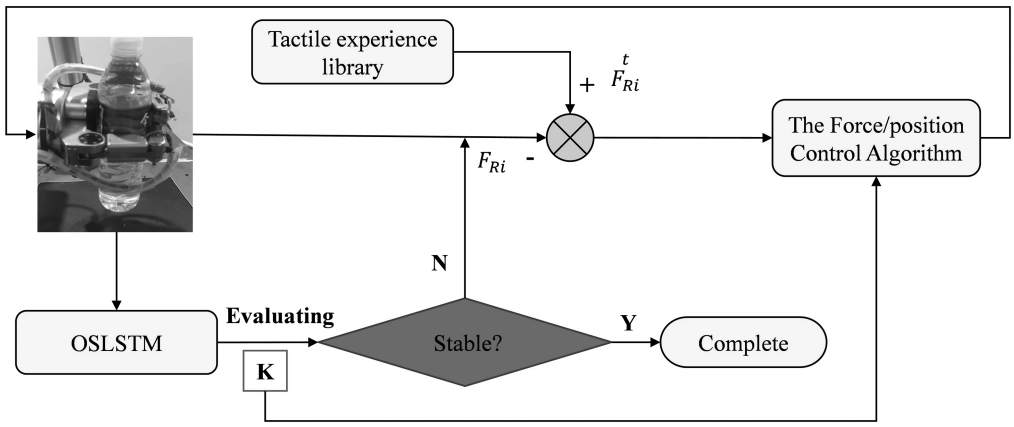

Figure 1. Pre-grasp process of visual-tactile fusion.

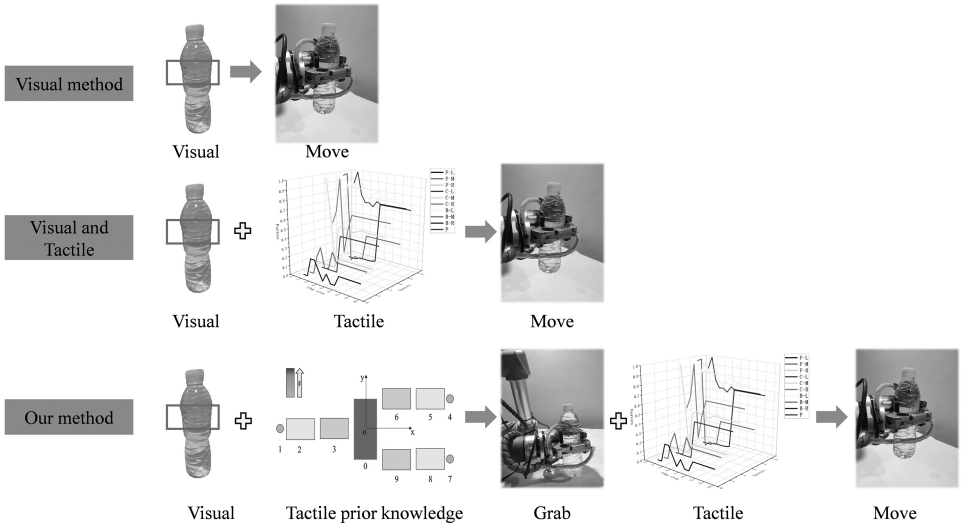

Figure 2. Examples of visual training samples.

corresponding to the distance between the opposite fingers before grasping, $Q(t,s) = \lambda_t \sum_n^{i=1} \ln(t_i+1) + \lambda_s \sum_n^{i=1} \ln(s_i+1)$ corresponding to the width of the contact position between the fingers and the target, $Q(t,s) = \lambda_t \sum_n^{i=1} \ln(t_i+1) + \lambda_s \sum_n^{i=1} \ln(s_i+1)$ corresponding to the direction of the grasp rectangle relative to the horizontal axis, and $Q(t,s) = \lambda_t \sum_n^{i=1} \ln(t_i+1) + \lambda_s \sum_n^{i=1} \ln(s_i+1)$ corresponding to one of the six categories corresponding to the target. The examples of visual labels are shown in Fig. 3.

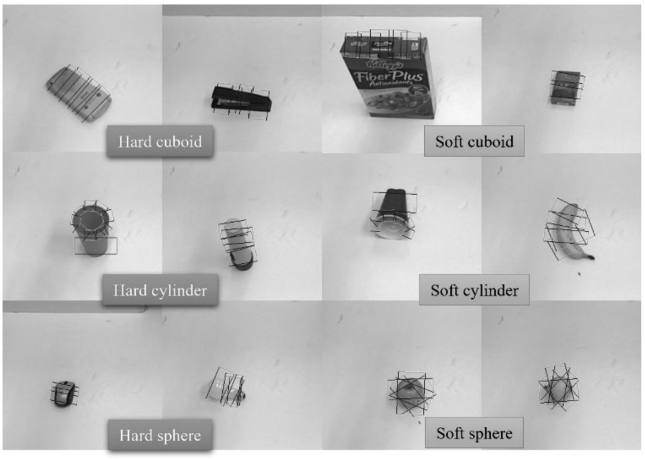

Figure 3. Examples of visual labels.

2.1.2 Touch: a Grasp Quality Evaluation Method and Grasp Stiffness Were Proposed

The distribution of the tactile sensors is shown in Fig. 4. There are 10 tactile sensors, which are respectively located: left fingertip (F-L), middle fingertip (F-M), right fingertip (F-R), left finger center (C-L), middle finger center (C-M), right finger center (C-R), left finger bottom (B-L), middle finger bottom (B-M), right finger bottom (B-R), and palm (P).

For obtaining the tactile sample data, we select four representative objects from each class of visual images. Then, we control the robot to repeatedly grasp each object and recorded the tactile data in the process of grasping.

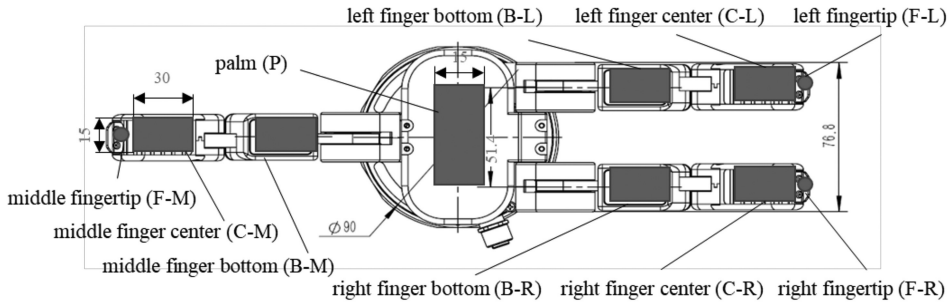

Figure 4. The distribution of tactile sensors.

We use (1) to define the grasping state.

$$Q(t,s) = \lambda_t \sum_{i=1}^{n} \ln(t_i + 1) + \lambda_s \sum_{i=1}^{n} \ln(s_i + 1) \qquad (1)$$

Where n is the number of fingers, t_i and s_i are the tactile average reading and motor current reading of the i-th finger respectively. λ_t and λ_s are the scale factors of tactile and strain perception, respectively, which are used to adjust the weight of each mode. They satisfy the relation in (2).

$$\begin{aligned} \lambda_t &= \frac{1}{k_t \ln(t_m + 1)}, \\ \lambda_s &= \frac{1}{k_s \ln(s_m + 1)}, \qquad \mathrm{n}\left(\frac{1}{k_t} + \frac{1}{k_s}\right) = 1 \end{aligned} \qquad (2)$$

Where, t_m and s_m are the maximum value of tactile value and current value when grasping, respectively, while k_t and k_s are the weights.

The Q value represents the gripping tightness of the robot. The higher the Q value, the tighter the grip. Due to the different materials and shapes of different objects, the gripping tightness varies greatly. For example, the Q value of the soft material is small but this does not mean that the clamping fails. Therefore, we use the Q-value difference to represent the degree of change of the robot's grasping state. The larger the Q-value difference, the more unstable the robot grasps and the greater the possibility of failure. When the Q-value difference is less than the set threshold, the grasping is considered stable. The calculation formula is shown in (3).

$$|Q_1 - Q_2| < \eta \qquad (3)$$

During the tactile data collection process, the robot is taught to grasp the position and the grasping pattern by humans. The robot is then asked to repeat the teaching actions to obtain more data. Firstly, the robot is taken a firm grasp on the object by recording the grasping state value $Q1$. The fingers are then lifted and swung to test the state. This is to verify the stability of the robot on the one hand and to collect more failure samples on the other hand. If the object fell or slipped, it would be judged as a failure. If not, the grasping state value $Q2$ would be recalculated after the swing process. The grasp is marked as successful when the difference between the two evaluations is less than the threshold value (η is set to 1.0) and there is no damage on the object's surface. Otherwise, it would be recorded as a failure. Combined with the vulnerability of the target and the output current of the finger in the successful grasping experiment, the grasp hardness is defined for each type of object as shown in Table 1. There are a total of 12,000 tactile data sets collected, including 8,400 successful samples and 3,600 unsuccessful samples (the collected visual samples are only used to test the visual grasping network, not as a training set).

The tactile data are shown in Fig. 5. The tactile data are the contact force values when the tactile sensor is in contact with the object, and the feedback period is 10 ms. Our tactile samples are composed of the data continuously collected by the tactile sensor array. Each row of data is collected by each sensor at the same time. And each column is collected by a single sensor at different times. There are two tactile labels, the grasp hardness K and the grasping outcome (1 for success and 0 for failure).

2.2 The Deep Learning Network for Grasp Position Detection

To realise the direct mapping from the image to the stable grasping instruction, this paper presents a visual-tactile fusion grasp planning model, which includes two stages: the grasp position detection and the grasp force distribution generation. The grasp position detection is completed by the CNN network as shown in Fig. 6. The network takes 512 × 512 × three color images as input, including six convolution layers, and two fully connected layers. Each convolution layer is followed by an activation function, a pooling layer. The last layer is a classification layer. The output of the network includes the coordinate (x, y) of the grasp center, w and h of the grasp area, θ of the grasp angle, the class C of the target, and the corresponding probability P. In order to associate the target category with the grasp box to get a more accurate grasp position, we have modified the loss function.

The loss function of the network output layer is divided into two parts, as shown in (4): the classification loss function for the category and the regression loss function for the position. The first part makes judgments on the target categories and uses a softmax classifier to judge the probability of each category, and the grasping rectangle box corresponding to the category with a low score will

Table 1
Examples of the Grasp Hardness

Shape/hardness	Object	Grasp hardness	Shape/hardness	Object	Grasp hardness
cuboid-hard	scissors	10	cuboid-soft	packing box	4
	calculator	8		book	6
	stapler	10		milk carton	4
	camera	7		remote control	6
cylinder-hard	ceramic cup	8	cylinder-soft	banana	2
	screwdriver	10		paper cup	1
	club	9		pop can	4
	glass cup	7		adhesive tape	4
sphere-hard	mouse	9	sphere-soft	apple	6
	ceramic bowl	7		pear	5
	potato	7		orange	3
	lamp	6		rubber ball	2

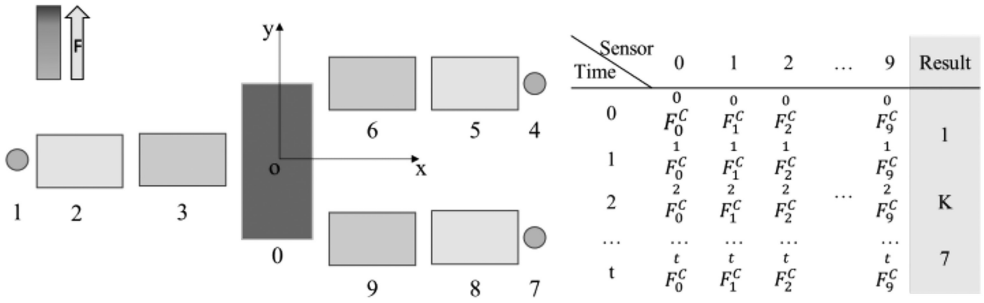

Figure 5. Examples of tactile samples.

be directly excluded. The second part is a regression loss function for the predicted geometric offset parameter of the grasping rectangular box to the reference rectangular box.

$$L(u,v) = -\lambda \log \frac{e^u}{\sum_{s=1}^{C} e^u s} + \eta \sum_{t \in \{x,y,w,h,\theta,n_c\}} \text{smoot } h_l\left(v^t - \widehat{v}^t\right) \quad (4)$$

Where, n_c is the number of categories, u is the softmax input corresponding to the correct class. v is the target location parameter obtained by network regression, $\widehat{v}$ is the actual target location parameter, λ is the coefficient of classification loss function, and η is the coefficient of the regression loss function. smooth_l is a smooth function of L1 loss function, as shown in (5).

$$\text{smoot } h_l(x) = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \quad (5)$$

Furthermore, the depth information is obtained by depth camera.

2.3 Pre-Grasp based on Tactile Prior Knowledge

As the convolutional neural network is unable to output the material of the object, it cannot determine how much grasp force the robot's finger should apply. If the gripper force was too strong, the object would be deformed and damaged. The insufficient forces may cause a slide. This paper establishes a tactile prior knowledge base to determine the pre-grasp force. We obtain a considerable amount of tactile data through the sampling process shown in Fig. 2. The distribution of the tactile data is shown in (6).

$$F_S = [F_{S1}, F_{S2}, \ldots, F_{S10}] \quad (6)$$

Then, we use (7) to establish the tactile experience value.

$$F_{\text{Si}}^{C} = \frac{\sum_{t=1}^{n} \overset{t}{F}_{\text{Si}}^{C}}{N} \quad (7)$$

Where N is the total number of successful samples, $\overset{t}{F}_{\text{Si}}^{C}$ is the force corresponding to the successful grasp of the target in the tth sampling, F_{Si}^{C} is the standard value of ith sensor for class C.

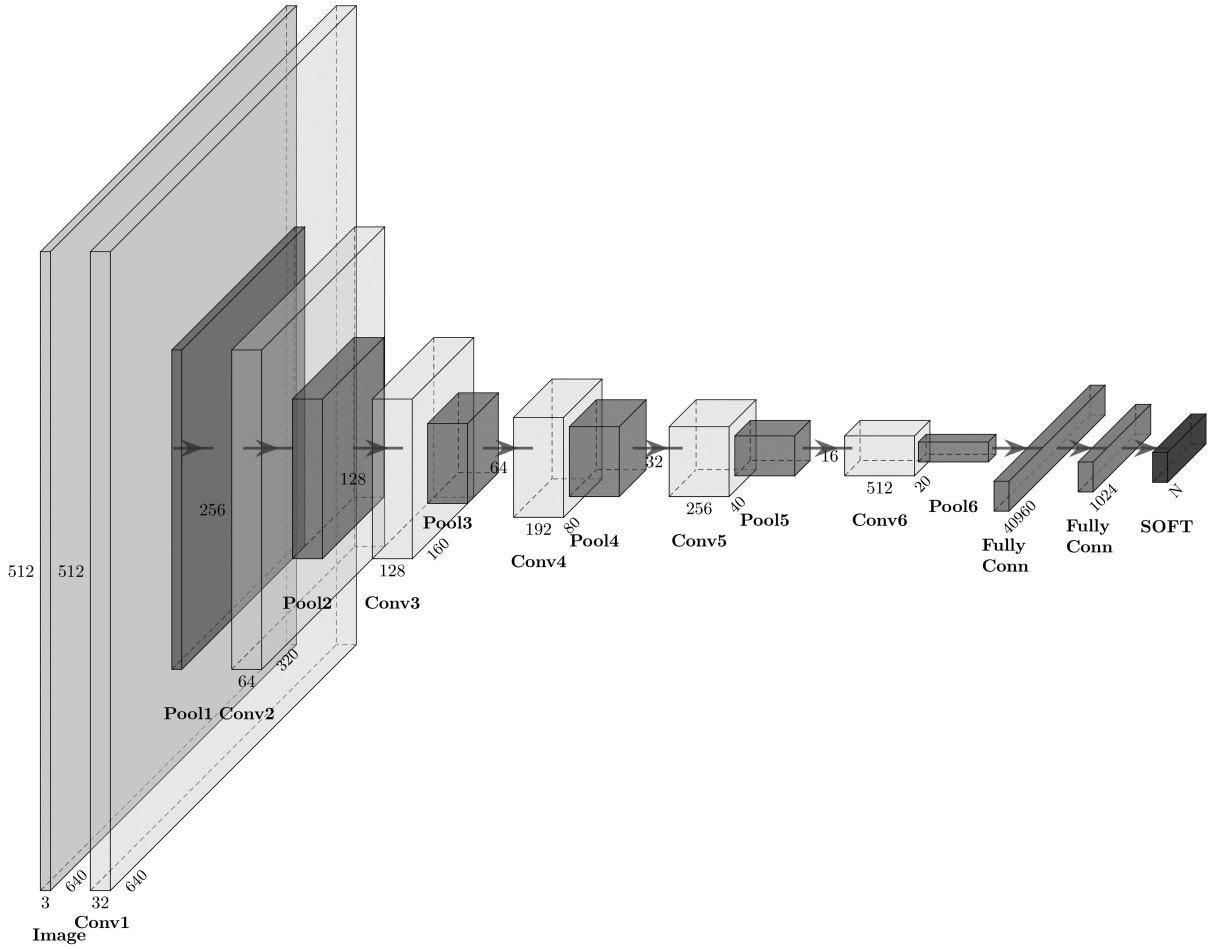

Figure 6. The network structure of grasp position detection.

Then, combined with the detection results of the deep learning network, we can get the pre-grasp force by (8).

$$F_{\mathrm{Ri}} = \sum_{C=1}^{N} F_{\mathrm{Si}}^{C} P_C, i = 1, 2 \ldots 10 \tag{8}$$

Where N is the total number of output categories, C is the output class of the detection network, P_C is the confidence, F_{Si}^{C} is the corresponding standard force value of the class, n is the number of all possible categories detected, and F_{Ri} is the pre-grasp force of the target.

Because the tactile experience value has been completed in the preparatory work, we can obtain all of the information about the pre-grasp task from a single visual image in the actual detection. Furthermore, we eliminate probable collisions and damage in the traditional tactile exploration process. By using tactile previous knowledge, we increase the safety and stability of grasping.

3. Robot Force/Position Control based on OSLSTM

3.1 OSLSTM

After successfully grasping the object, the robot will move the object to complete the assembly and other tasks. The grasping strategy should be adjusted in real time in response to changes of the object state while the robot is moving. In this paper, the improved OSLSTM neural network is used to identify the grasp hardness and the grasping state of the object. The network [32], [33] is a type of time recursive neural network that is suitable for predicting events with a relatively long interval in time series. While LSTM overcomes the problem of long-term reliance, it does have a limit [34].

The LSTM network takes a tactile sequence as input which has two dimensions. The horizontal dimension is that we collect data from 10 tactile sensors at the same time. The width is 10. The vertical dimension is the number of times we collect, which is the time step of the LSTM. It is variable as an important parameter of the LSTM. The training process for the model would generate noticeable oscillation and the training period would be protracted if the time step was too large. Furthermore, in the actual test, real-time performance is poor and gripping efficiency is low. On the contrary, if the time step was set too tiny, the continuity for the grasping process would be quickly lost, resulting in low accuracy. In the experiment, it was discovered that the ideal step size for different sorts of objects differed. Aiming at the problem of the length of time step in the LSTM, this paper designs a tactile OSLSTM network. Through the time step self-search algorithm, the optimal time step will be determined by the change of sensor data and adjusted in real time, allowing the network to achieve

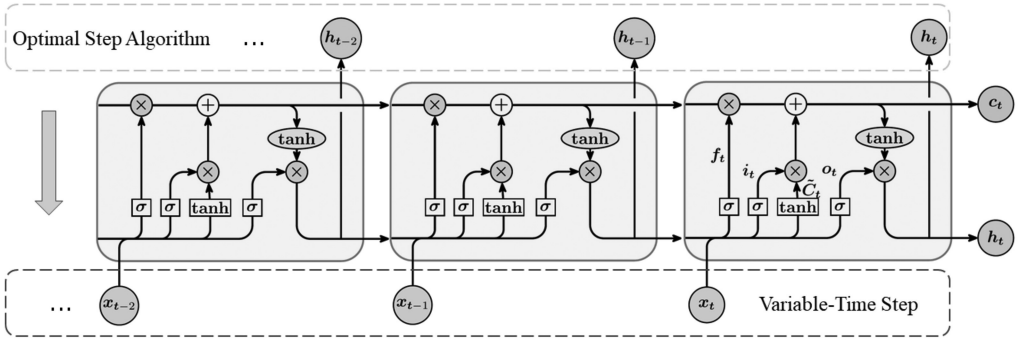

Figure 7. The network structure of the OSLSTM.

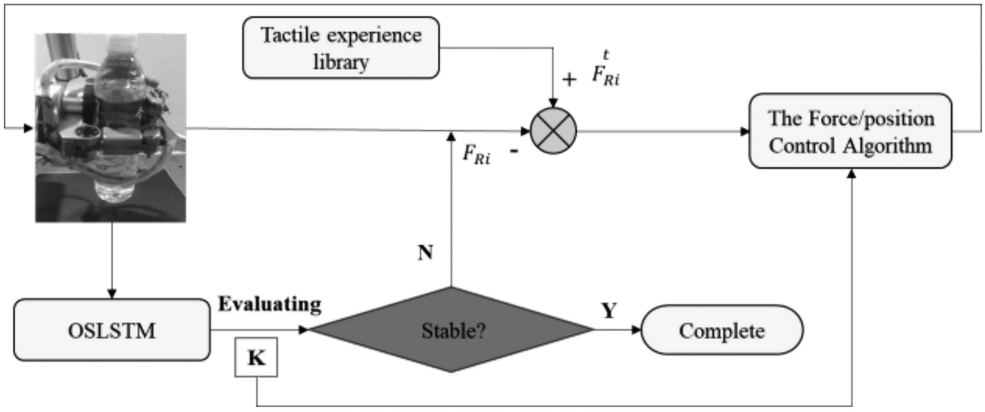

Figure 8. The force/position control process.

higher performance. The tactile network structure is shown in Fig. 7.

The optimal time step self-search algorithm is shown in (9) to (15). The search direction and increment of step size are determined according to the change of loss function. The algorithm iterates until the loss function fulfills the requirements. Then, the step size is the optimal step size.

$$x^{(k+1)} = x^{(k)} + \lambda_k d^{(k)} \tag{9}$$

Where $x^{(k)}$ is the time step of the kth search, λ_k is the step increment of the kth search, d is the search direction, the initial value: $x^{(1)} = 15, d = 1, \lambda = 5$. And the loss function is as shown in (10).

$$\varphi(x) = k_1 t(x) + k_2 \frac{1}{p(x)} \tag{10}$$

Where $\varphi(x)$ is the objective loss function, $t(x)$ is the function of time varying with the time step, $p(x)$ is the function of network accuracy varying with the time step. k_1 and k_2 are the weights.

The purpose of the algorithm is to find the time step that minimises the objective loss function. The iterative method is shown in (11).

$$d_{k+1} = \text{sign}\left(\varphi_k(x) - \varphi_{k-1}(x)\right) \tag{11}$$

Where, sign (α) is the value function, as shown in (12).

$$\text{sign}(\alpha) = \begin{cases} 1, \ \alpha \geq 0 \\ -1, \ \alpha < 0 \end{cases} \tag{12}$$

$$\lambda_{k+1} = \begin{cases} \lambda_k, d_{k+1} = 1 \\ [\lambda_k/2], d_{k+1} = -1 \end{cases} \tag{13}$$

$$\varphi(x) < \varepsilon \tag{14}$$

$$x = \arg\min_x \varphi(x) \tag{15}$$

The iterative of λ value is shown in (13) and the threshold condition is shown in (14). The optimal step is shown in (15).

3.2 The Force/Position Control Algorithm based on the OSLSTM

After obtaining the grasp hardness and state of the target by the OSLSTM network, we can adjust the grasp force and position according to the force/position control algorithm when the unstable grasp state is detected. Firstly, the force sensors are used to obtain the deviation between the actual force and the expected force when the robot contacted the target. Secondly, the force deviation is converted into the position by the grasp hardness. Finally, the robot moves from its current position to the desired position to complete the force control tasks. The force/position control process is shown in Fig. 8.

The force position control algorithm is as (16) to (19).

$$X = (x_1, x_2, \ldots, x_{10}) \tag{16}$$

$$X_R = (x_{r1}, x_{r2}, x_{r3}) \tag{17}$$

$$X = k \times K e^{(F_{\mathrm{Ri}} - \overset{t}{F}_{\mathrm{Ri}})} \tag{18}$$

$$\begin{cases} x_{r1} = \frac{x_1+x_2+x_3}{3} \\ x_{r2} = \frac{x_4+x_5+x_6}{3} \\ x_{r3} = \frac{x_7+x_8+x_9}{3} \end{cases} \tag{19}$$

Where, X is the adjustment required for each tactile sensor point, and X_R is the offset required for each finger, $\overset{t}{F}_{\mathrm{Ri}}$ is the current force of tactile sensors, and F_{Ri} is the empirical force value in the stable state. K is the grasp hardness, which has different values for different categories. We measure the grasping hardness of only some of the objects in our experiments and then input them to the network for training to establish the relationship between haptic distribution and grasping stiffness. k is the transformation ratio between position and grasping force in mm/N, which is 1.5 as measured by the experiments.

As each finger of three is driven by a motor, we get the corresponding offset of each finger from (19). We are able to obtain the movements of the fingers based on the object's grasp hardness using the OSLSTM network. The robot is adjusted constantly until the network's output condition is stable.

4. Experiments

The experiment platform included a three-fingered robot hand, array of tactile sensors, a Kinect2 camera, and the UR5 robot arm, as shown in Fig. 9. The UR5 robot arm stands at the left of the table, upon which we assemble the three-fingered robot hand with 10 tactile sensors. The dexterous hand is the JQ3-5 of the JODELL company. The tactile sensors utilised in this research are flexible thin-film piezoresistive sensors with an impressive response time of less than 1ms and a data feedback period of less than 5 ms. A Kinect2 camera is at the right of the scene to record the view of grasping. The PC was configured with a 2.7 GHz frequency and16 GB of memory. It was equipped with an NVIDIA GeForce2080Ti with a computing capability of 7.5.

We build a visual and tactile grasp dataset that included a variety of characteristics, such as shape, size, weight, and grasp style. The visual samples included the Cornell Grasp Dataset and our own objects images. The objects in the dataset contain different sizes of cuboids, cylinders, and spheres.

4.1 Grasp Position Detection Experiment

We evaluate our approach on the Cornell Grasp Dataset and utilised five-fold cross validation for our experimental data to compare it to other gripping methods. For 48 objects we have a total of about 200 visual samples. The dataset is split in two different ways:

1) Image-wise split: Splits images randomly.

Figure 9. The experiment platform.

2) Object-wise split: Object-wise splitting randomly separates all object instances and all images of an object are grouped in a single validation set. This is helpful to test how well did the network generalise to novel objects.

As the rectangle metric is better at discriminating between "good" and "bad" grasping positions, we use this metric for our experiments. The rectangle metric considers a grasp to becorrect if both:

1) The grasp angle is within 30° of the ground truth grasp.

2) The Jaccard index of the predicted grasp and the ground truth is greater than 25 percent.

Where the Jaccard index is given by (20). The formula shows the overlap between the two regions, which is an important indicator to evaluate the quality of the detection.

$$J(\widehat{S}, S) = \frac{|\widehat{S} \cap S|}{|\widehat{S} \cup S|} \tag{20}$$

where $\widehat{S}$ is the predicted grasp rectangular boxes and S is the ground truth grasp rectangular boxes.

The experimental results are shown in Table 2. The experiment is divided into two steps: detection and grasping. The grasping experiment is to grasp the target after successfully detecting it. In this paper, the success rate of the non-destructive grasp is used to evaluate the quality of initial grasping. The initial grasp should ensure that there are no scratches or destructive deformations on the object's surface. Otherwise, the grasp is failed. This paper incorporates the tactile prior knowledge after detection. While other detection methods use a set uniform size force of 10 N because of no tactile experience. Therefore, in the grasping experiments, other detection methods are used as a group to compare with this paper's method. This group of experiments is only to verify that the inclusion of tactile information can improve the quality of grasping. For each type of object, we performed 10 sets of grasping tests. A total of 480 sets of grasping experiments were performed.

We categorise objects into six groups based on their shapes and materials. These characteristics are linked to the method by which they are grasped. According to the experiments, our method obtains an accuracy of 85.9% for image-wise split and 88.3% for object-wise split, which is slightly poorer than other methods. However, we especially outperform in the object-wise split. Although we do not

Table 2
Comparison of Object Detection Algorithms

Detection algorithm	Accuracy			Time/ms
	Image-wise split	Object-wise split	Non-destructive grasp (after successful detection)	
Jiang *et al.* [35]	0.605	0.583	0.754	5000
Lenz *et al.* [4]	0.739	0.756		1350
Redmon and Angelova [36]	0.880	0.871		76
Kumra and Kanan [37]	0.848	0.845		103
Morrison *et al.* [38]	0.730	0.690		19
Karaoguz and Jensfelt [39]	0.887	/		200
OURS	0.859	**0.883**	**0.912**	**6**

directly include tactile information in the visual training, we classify the targets into six categories based on the material and shape properties of the objects, and such classification is closely related to the tactile features. After recognising these attributes of the target, the neural network can quickly determine the best grasping box. We don't perform particularly well in terms of visual inspection performance, but our structure is simple and our computation time is the fastest.

In addition, while the performance of our method is worse than the methods of the Redmon and Karaoguz in image-wise split, our method has a significant improvement in the success rate of non-destructive grasp. For some very fragile objects, damage has been caused during the initial grasp. Consequently, the non-destructive grasp's success rate was still only 75.4%. In our method, we added the fusion of visual detection and tactile prior knowledge based on visual-tactile fusion. According to the visual detection results, the pre-grasp force was obtained through the established tactile prior knowledge, which approached the force value of stable grasp in advance. To a certain extent, damage caused by improper force setting was avoided. Finally, the success rate of the non-destructive grasp was increased to 91.2%. The experiment showed that our method could effectively detect the grasp position.

The core of the algorithm in this paper lies in the integration of visual information and tactile prior knowledge to improve the success rate of damage-free grasping. And there was no in-depth study and comparison of visual detection networks. Actually, the visual network in this method could be replaced by the currently available neural networks with better results.

4.2 Performance Tests of the OSLSTM Network

4.2.1 Optimal Self-Search of the Time Step

In the experiment, different shape samples were used to test LSTM performance under different time steps, and the influence of different shape samples on the network was analysed. The training loss of the network for different shapes at different time steps is shown in Fig. 10. Under the condition of the same training sample size, the network training time of different time steps is shown in Fig. 11. Then, according to the experimental results, three groups of proper network steps were compared with the OSLSTM algorithm. The impact of the fixed step and self-search step on network performance was analysed. The experimental results are shown in Table 3. The training sample size is 9,600 and the testing sample size is 2,400.

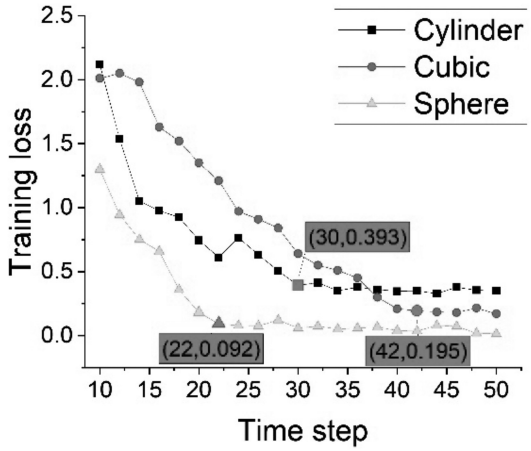

Figure 10. The training loss of the LSTM for different shapes at different time steps.

As shown in Fig. 10, under varied shape samples, the effect trend of step size on network performance is essentially the same. With the increase in step length, the training loss is smaller, but the training time is also increased. And after the loss had reached a certain level, increasing the step size has no obvious effect on the network loss. It indicates that there is overfitting. However, as shown in Fig. 11, the training time rises dramatically. The longer the sequence length is, the more noise data may be contained, which will lead to low accuracy. It is worth noting that the critical value of the step size of different shapes is not the same, corresponding to 22, 30, and 42 respectively.

Table 3
The Results of the LSTM Algorithm Comparison Experiment for All Samples

Groups	Time step	Training time (h)	Training loss	Testing time (ms)	Testing loss	Accuracy
LSTM-22	22	3.1	0.601	26.2	0.748	0.727
LSTM-30	30	6.0	0.364	43.8	0.591	0.850
LSTM-42	42	25	0.122	70.4	0.153	0.941
OSLSTM	15–35	3.5	**0.109**	**38.7**	**0.127**	**0.965**

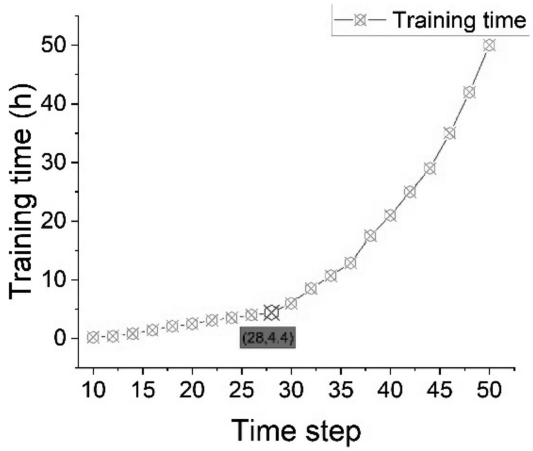

Figure 11. The training time of the LSTM for different time steps.

Considering the previous experimental foundation and the complexity of the actual verification operation, we selected steps 22, 30, and 42 as the LSTM steps to train and test the total sample, and compared them to the self-search algorithm created in our method. The experimental results are shown in Table 3. The OSLSTM network differed from the standard LSTM network in that the network step is optimised based on the detection samples. The experimental results indicate that the network step in our method changed in the range of 15–35. Compared with the fixed step, our algorithm obtains higher accuracy when the average step size is lower. Compared with the first group of experiments, the average step size of our algorithm is larger, but the training time does not increase significantly. Compared with the third group of experiments, in the case of a small average step, the training loss and accuracy are better. Although the greater the step size for a single sample, the stronger the effect, the optimal step size of each type of sample is not the same. When the step size is too large in the total sample, the model training produces an obvious oscillation phenomenon, resulting in a decrease in accuracy. Therefore, to improve network performance, the LSTM network is required to find the appropriate step size based on the target's properties.

4.2.2 Recognition of the OSLSTM Network

In the experiment, it is also discovered that there are some variations in the contact area and the effective unit of the tactile sensor array in the process of grasping the target for different objects. When grasping a cylindrical object, the fingertip will not touch the object. The data of the three sensors F-L, F-M, and F-R on the fingertip are all zero, as shown in (a1) and (a2) in Fig. 12. On the contrary, when grasping a square object, the robot relies heavily on its fingertips for grasping. As shown in (b1) and (b2), only the values of the fingertip sensors change. For spherical objects, the contact area is greater and almost every sensor has a numerical change.

Generally, for objects of the same shape with different materials, because the contact area of the dexterous hand between the rigid object is smaller than that between the soft object, the number of effective units of the tactile sensor array is less and the tactile data is sparser than that of soft objects. (c1) and (c2) in Fig. 12 show that the dexterous hand is used to grasp the hard ball and the soft leather ball respectively. It can be seen from the (c1) that when the dexterous hand grasps the hard ball, the effective unit number of the tactile sensor array is less. While its tactile feedback force is large. And the tactile data rapidly reaches its highest value. On the contrary, for the soft ball shown in (c2), the tactile sensor array has more effective units and less tactile feedback force. And the soft ball deforms progressively under the dexterous hand's pressure, resulting in a steady increase in feedback force that finally reaches its highest value. And the number of tactile sensors contacted also increases gradually.

In different grasping stages of the same object, the values and distribution of tactile data are inconsistent, as shown in (a) and (b) in Fig. 13(a) shows the moving process of the robot. Its tactile data is unevenly distributed and the values are smaller. It is a chaotic noise signal. This is frequently followed by the phenomenon of slipping in the experiment. In the stable state of grasping shown in Fig. B, there are fewer fluctuations among tactile sensors' data and the data values are bigger. The OSLSTM network can predict the different states of the same object based on these differences, then determines the stable grasping experience value based on the current movement attributes, and ultimately adjusts the gripper by the force/position control algorithm to assure the stability for grasping.

4.3 Visual-Tactile Fusion Grasping Experiment

Three groups were designed, including the vision, traditional visual-tactile fusion, and the proposed in this study. We controlled the robot to accelerate and rotate after

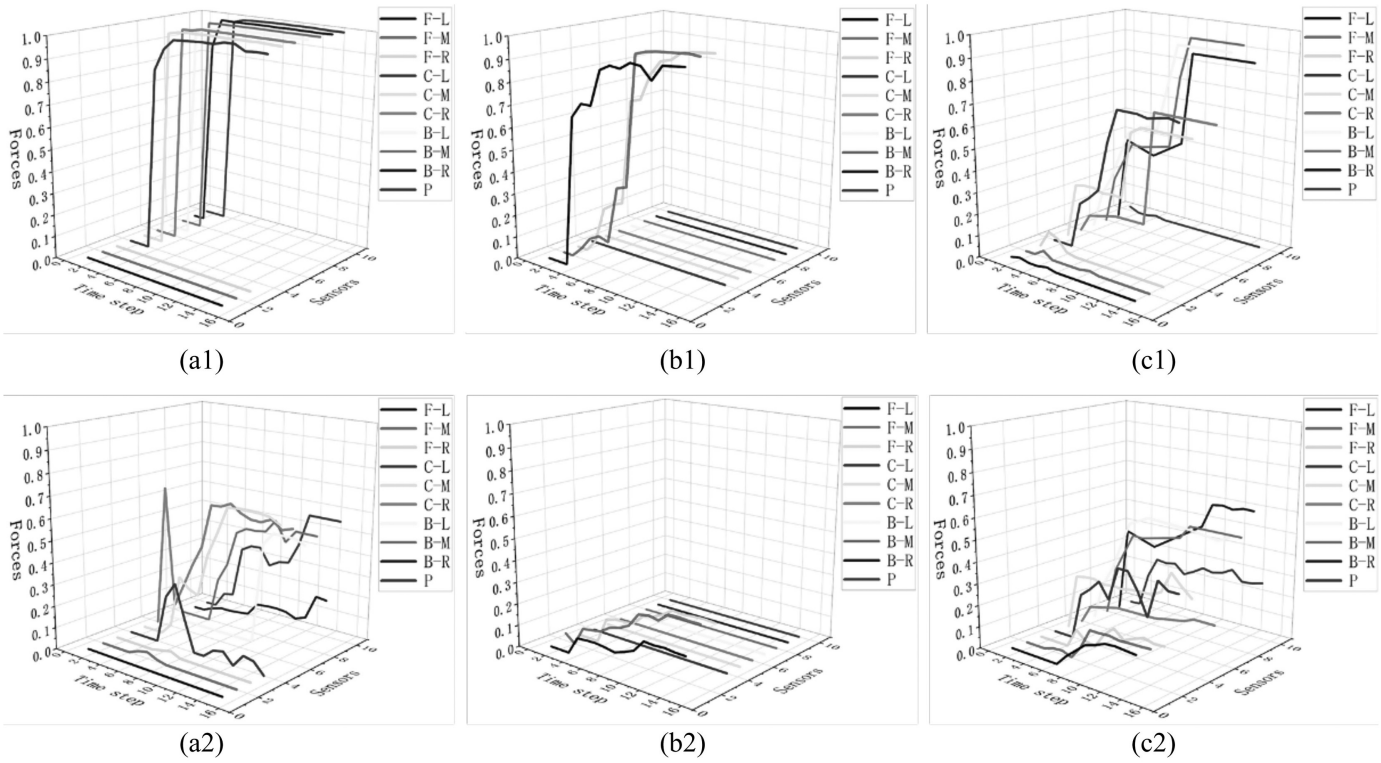

Figure 12. Pressure trend of tactile sensors under different objects: (a1) hard cuboid; (a2) soft cuboid; (b1) hard cylinder; (b2) soft cylinder; (c1) hard sphere; and (c2) soft sphere.

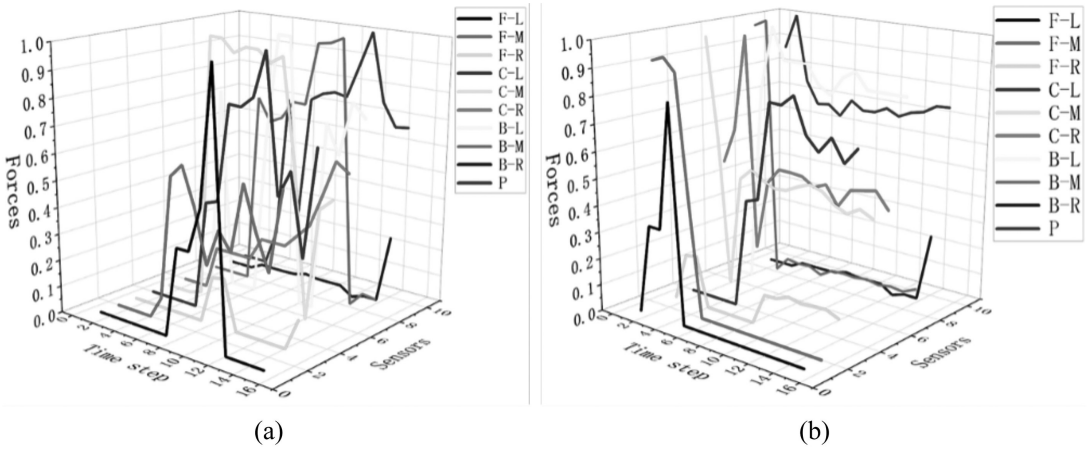

Figure 13. Pressure trend of the tactile sensors in different grasping stages of the same object: (a) unsteady state; (b) steady state.

grasping objects for testing the stability in each group. The detailed flow is shown in Fig. 14. The experimental results are shown in Table 4.

Vision: The visual detection algorithm was used to get the grasping posture. The robot grasped without tactile sense.

Fusion of Unadjusted Grasping: First, the target category and grasping posture were obtained by visual detection. Then, the object was grasped by the robot according to the pre-grasp force.

Fusion of Adjusted Grasping: The previous steps are the same as the fusion of unadjusted grasping. However, during the grasping process, the robot was regulated by the OSLSTM.

For each type of object, we performed 20 sets of grasping tests, and each set of experiments contained 540 grasping results. For the non-invasive grasping experiments, both unadjusted grasping and adjusted grasping were used based on the tactile prior knowledge of grasping. The grasping results were combined and calculated. In the mobile grasping process, both the visual method and the unconditioned method do not have a conditioning mechanism. Similarly, the grasping results were combined and computed.

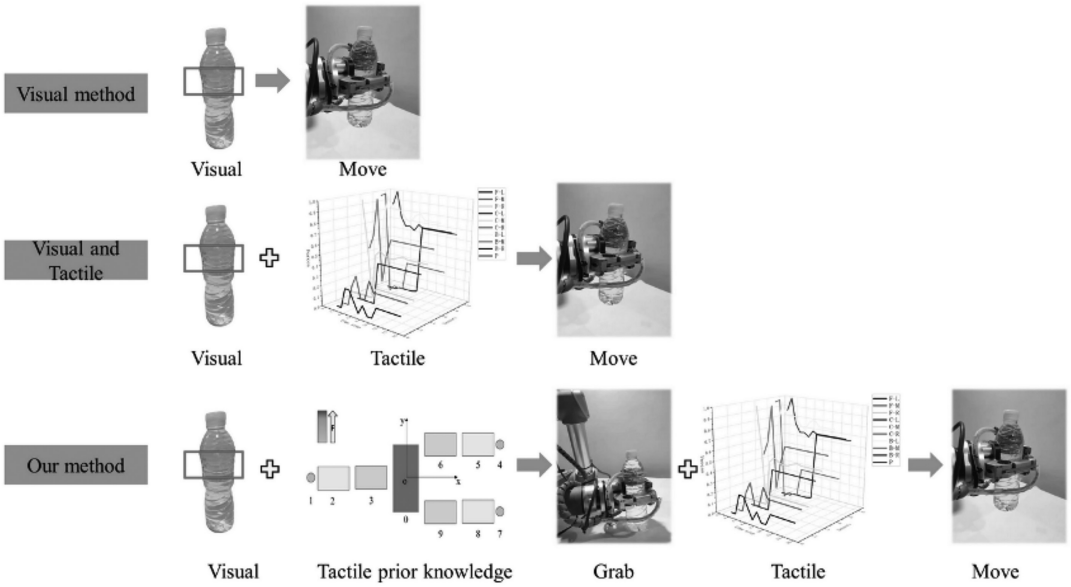

Figure 14. The process of different grasping strategies.

Table 4
Comparison of Different Grasping Methods

Algorithm	**The success rate of non-destructive grasp**	**The success rate of mobile grasp (after successful grasp)**
Vision	0.713	0.759
Fusion of unadjusted grasping	0.822	0.757
Fusion of adjusted grasping	**0.828**	**0.930**

The process of the dexterous hand successfully grasping the target object was divided into two stages: grasping the object without damage and grasping the object stably in the moving process. First, the experiment further confirms that the grasping method with the tactile experience database can reduce the damage to the object compared with direct grasping. In addition, the mobile grasping success rate of the first and second group methods without force/position control is only 75.9% and 75.7%, respectively. With the object in motion, the grasping position changes. The vision can not perceive the force change, resulting in the object sliding. The third group of experiments increases the mobile grasping success rate to 93.0% after adding the force/position control algorithm. By detecting the grasping state of the object, the robot can change the grasping strategy in real time until the grasping is stable. In the experiments, this paper finds that the grasping parameters of the dexterous hand change when the pose and appearance of the object change. Therefore, only by detecting the grasping state in real time, a more reasonable robot finger displacement can be calculated.

5. Conclusion

The use of the integration of visual detection and tactile perception to robotic grasp and manipulation tasks is discussed in this study. We discover a link between visual and tactile information. Safe and effective pre-grasping are achieved by the combination of visual identification and tactile prior knowledge. In addition, we build detecting network of the grasp position and enhance the LSTM network's performance. We address the difficulty of the unstable grasp phenomena caused by the robot's movement by combining the grasp hardness and the force/position control algorithm, laying the foundation for the robot's compliant manipulation.

However, the pre-grasp network and the force/position control algorithm are heavily dependent on tactile experience, necessitating more tactile sample data gathering. It requires a huge number of training samples. Furthermore, this paper does not deeply study the potential relationship between visual data and tactile data. In fact, the association between visual and haptic information is much closer than we know. In the future, fusion at the data level can be further investigated to analyse its impact on robot manipulation. For example, a new fusion network of visual and tactile data could be created, taking raw visual and tactile information as input and outputting corresponding robot actions to establish an end-to-end relationship between fused data and robot control. The haptic sensor can also be improved to increase the number of tactile sensing units, thus obtaining richer tactile information.

References

[1] Y. Liu, Z. Li, H. Liu, Z. Kan, and B. Xu, Bioinspiredembodiment for intelligent sensing and dexterity in fine manipulation: A survey, *IEEE Transactions on Industrial Informatics, 16*(7), 2020, 4308–4321.

[2] V. Vellaiyan, S. Subramaniam, and V. Arunachalam, Bend angle and contact force on soft pneumatic gripper for grasping cylindrical-shaped different-sized objects, *International Journal of Robotics and Automation, 37*(5), 2022, 391–399.

[3] N. Korbinian, S. Arne, and A.-S. Alin, Towardsautonomous robotic assembly: Using combined visual and tactile sensing for adaptive task execution, *Journal of Intelligent & Robotic Systems, 101*(3), 2021.

[4] I. Lenz, H. Lee, and A. Saxena, Deep learning for detecting robotic grasps, *The International Journal of Robotics Research, 34*(4–5), 2015, 705–724.

[5] L. Pinto and A. Gupta, Supersizing self-supervision: Learning to grasp from 50Ktries and 700 robot hours, *Proc.2016 IEEE International Conf.on Robotics And Automation (ICRA)*, Stockholm, 2016, 3406–3413.

[6] E. Lyu, X. Yang, W. Liu, J. Wang, S. Song, and Q.-H.M. Max, AN autonomous eye-in-hand robotic system for picking objects in a supermarket environment with non-holonomic constraint, *International Journal of Robotics and Automation, 37*(4), 2022, 352–361.

[7] L. Chen, P. Huang, and Z. Meng, Convolutional multi-grasp detection using grasp path for RGBD images, *Robotics and Autonomous Systems, 113*, 2019, 94–103.

[8] A. Garcia-Garcia,S. Orts-Escolano, S.Oprea, J. Garcia-Rodriguez, J.Azorin-Lopez, M.Saval-Calvo, and M.Cazorla, Multi-sensor 3D object dataset for object recognition with full pose estimation, *Neural Computing and Applications, 28*(5), 2017, 941–952.

[9] R. Calandra, A. Owens, M.Upadhyaya,W. Yuan,J. Lin,E.H. Adelson, and,S. Levine, The feeling of success: Does touch sensing help predict grasp outcomes?, 2017, *arXiv:1710.05512*.

[10] R. Calandra, A. Owens, M.Upadhyaya,W. Yuan,J. Lin,E.H. Adelson, and,S. Levine, More than a feeling: Learning to grasp and regrasp using vision and touch, *IEEE Robotics and Automation Letters, 3*(4), 2018, 3300–3307.

[11] R. Li and E.H. Adelson, Sensing and recognizing surface textures using a GelSightsensor, *Proc.of the IEEE Conf.on Computer Vision and Pattern Recognition*, Portland, OR, 2013, 1241–1247.

[12] H. Hai, D.-P.Yang, C.-Y.Sun, N. Li, Y.-J. Pang, L. Jiang, and H. Liu, Surface EMG for multi-pattern recognition with sensory feedback controller of hand prosthesis system, *International Journal of Robotics and Automation, 28*(1), 2013.

[13] Y. Chebotar, K. Hausman, Z. Su, G.S. Sukhatme, and S. Schaal, Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning, *Proc.2016 IEEE/RSJ International Conf.on Intelligent Robots and Systems (IROS)*, Daejeon, 2016, 1960–1966.

[14] Z. Yi, T. Xu, W. Shang, W. Li, and X. Wu, Genetic algorithm-based ensemble hybrid sparse ELM for grasp stability recognition with multimodal tactile signals, *IEEE Transactions on Industrial Electronics, 70*(3), 2022, 2790–2799.

[15] Z. Deng, Y. Jonetzko, L. Zhang, and J. Zhang, Grasping force control of multi-fingered robotic hands through tactile sensing for object stabilization, *Sensors, 20*(4), 2020, 1050.

[16] Q. Li, O. Kroemer, Z. Su, F.F. Veiga, M. Kaboli, and H.J. Ritter, A review of tactile information: Perception and action through touch, *IEEE Transactions on Robotics, 36*(6), 2020, 1619–1634.

[17] C. De Farias, N. Marturi, R. Stolkin, and Y. Bekiroglu, Simultaneous tactile exploration and grasp refinement for unknown objects, *IEEE Robotics and Automation Letters, 6*(2), 2021, 3349–3356.

[18] F. Veiga, H.V. Hoof, J. Peters, and T. Hermans, Stabilizing novel objects by learning to predict tactile slip, *Proc.2015 IEEE/RSJ International Conf.on Intelligent Robots and Systems (IROS)*, Hamburg, 2015, 5065–5072.

[19] D. Han, H. Nie, J. Chen, M. Chen, Z. Deng, and J. Zhang, Multi-modal haptic image recognition based on deep learning, *Sensor Review, 38*(4), 2018, 486–493.

[20] T. Li, X. Sun, X.Shu, C. Wang, Y. Wang, G. Chen, and N.Xue, Robot grasping system and grasp stability prediction based on flexible tactile sensor array, *Machines, 9*(6), 2021, 119.

[21] J.W. James and N.F. Lepora, Slip detection for grasp stabilization with a multifingered tactile robot hand, *IEEE Transactions on Robotics, 37*(2), 2020, 506–519.

[22] S.J. Lederman and R.L. Klatzky, Multisensory texture perception,in J. Kaiser and M. Naumer(eds.), *Handbook of multisensory processes* (New York, NY: Springer, 2004), 107–122.

[23] S. Lacey, C. Campbell, and K. Sathian, Vision and touch: Multiple or multisensory representations of objects?, *Perception, 36*(10), 2007, 1513–1521.

[24] F. Wallhoff, J. Blume, A. Bannat, W. Rösel, C. Lenz, and A. Knoll, A skill-based approach towards hybrid assembly, *Advanced Engineering Informatics, 24*(3), 2010, 329–339.

[25] S. Wang, J. Wu, X. Sun, W. Yuan, W.T. Freeman, J.B. Tenenbaum, and E.H. Adelson, 3Dshape perception from monocular vision, touch, and shape priors, *Proc.2018 IEEE/RSJ International Conf.on Intelligent Robots and Systems (IROS)*, Madrid, 2018, 1606–1613.

[26] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, A hybrid deep architecture for robotic grasp detection, *Proc.2017 IEEE International Conf.on Robotics and Automation (ICRA)*, Singapore, 2017, 1609–1614.

[27] S. Cui, R. Wang, J. Wei, J. Hu, and S. Wang, Self-attention based visual-tactile fusion learning for predicting grasp outcomes, *IEEE Robotics and Automation Letters, 5*(4), 2020, 5827–5834.

[28] A. Caporali, K. Galassi, G. Laudante, G. Palli, and S. Pirozzi, Combining vision and tactile data for cable grasping, *Proc.2021 IEEE/ASME International Conf.on Advanced Intelligent Mechatronics (AIM)*, Delft, 2021, 436–441.

[29] Y. Han, K. Yu, R. Batra, N. Boyd, C. Mehta, T. Zhao, Y. She, S. Hutchinson, and Y. Zhao, Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer, 2021, *arXiv:2112.06374*.

[30] S. Kanitkar, H. Jiang, and W. Yuanl, PoseIt: A visual-tactile dataset of holding poses for grasp stability analysis, *Proc.2022 IEEE/RSJ International Conf.on Intelligent Robots and Systems (IROS)*, 2022, 71–78.

[31] M. Matak and T. Hermans, Planning visual-tactile precision grasps via complementary use of vision and touch, *IEEE Robotics and Automation Letters*, *8*(2), 2022, 768–775.

[32] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Computation, 9*(8), 1997, 1735–1780.

[33] A. Graves and J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks, 18*(5–6), 2005, 602–610.

[34] K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, LSTM: A search space odyssey, *IEEE Transactions on Neural Networks and Learning Systems, 28*(10), 2016, 2222–2232.

[35] Y. Jiang, S. Moseson, and A. Saxena, Efficient grasping from rgbd images: Learning using a new rectangle representation, *Proc.2011 IEEE International Conf.on Robotics and Automation*, Shanghai, 2011, 3304–3311.

[36] J. Redmon and A. Angelova, Real-time grasp detection using convolutional neural networks, *Proc.2015 IEEE International Conf.on Robotics and Automation (ICRA)*, Seattle, WA, 2015, 1316–1322.

[37] S. Kumra and C. Kanan, Robotic grasp detection using deep convolutional neural networks, *Proc.2017 IEEE/RSJ International Conf.on Intelligent Robots and Systems (IROS)*, Vancouver, BC, 2017, 769–776.

[38] D. Morrison, P. Corke, and J. Leitner, Learning robust, real-time, reactive robotic grasping, *The International Journal of Robotics Research, 39*(2–3), 2020, 183–201.

[39] H. Karaoguz and P. Jensfelt, Object detection approach for robot grasp detection, *Proc.2019 International Conf.on Robotics and Automation (ICRA)*, Montreal, QC, 2019, 4953–4959.

Biographies

Zihao Ding was born in 1996. He received the B.Sc. degree from Wuhan Textile University, China, in 2017, He is currently pursuing the Ph.D. degree with the School of Mechanical and Electric Engineering, Soochow University. His research interests include robot vision and robotics.

Guodong Chen was born in 1983. He received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2011. He is an Associate Professor with Soochow University. His research interests include robot vision and intelligent industrial robot.

Zhenhua Wang was born in 1974. He received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China. His research interests include Industrial robots and intelligent automation equipment. He is young and middle-aged academic leaders of the "Blue Project" in Jiangsu Province.

Lining Sun was born in 1964. He received the Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 1993. He is currently a Professor with Soochow University. His research interests include industrial robot and intelligent manufacturing.