International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 1753–1762
DOI: https://doi.org/10.2991/ijcis.d.210531.001; ISSN: 1875-6883; eISSN: 1875-6891
https://www.atlantis-press.com/journals/ijcis/

Research Article

Tactile–Visual Fusion Based Robotic Grasp Detection Method with a Reproducible Sensor

Yaoxian Song[1,2], Yun Luo[2], Changbin Yu[3,4,*]

[1]*School of Computer Science & Institute for Intelligent Robots, Fudan University, Shanghai, China*
[2]*School of Engineering, Westlake University, Hangzhou, China*
[3]*College of Artificial Intelligence and Big Data, Shandong First Medical University & Shandong Academy of Medical Sciences, Shandong, China*
[4]*Faculty of Engineering & the Built Environment, University of Johannesburg, Johannesburg, South Africa*

ARTICLE INFO

Article History
Received 15 Mar 2021
Accepted 26 May 2021

Keywords
Tactile sensor
Tactile–visual dataset
Multi-modal fusion
Deep learning
Grasp detection

ABSTRACT

Robotic grasp detection is a fundamental problem in robotic manipulation. The conventional grasp methods, using vision information only, can cause potential damage in force-sensitive tasks. In this paper, we propose a tactile–visual based method using a reproducible sensor to realize a fine-grained and haptic grasping. Although there exist several tactile-based methods, they require expensive custom sensors in coordination with their specific datasets. In order to overcome the limitations, we introduce a low-cost and reproducible tactile fingertip and build a general tactile–visual fusion grasp dataset including 5,110 grasping trials. We further propose a hierarchical encoder–decoder neural network to predict grasp points and force in an end-to-end manner. Then comparisons of our method with the state-of-the-art methods in the benchmark are shown both in vision-based and tactile–visual fusion schemes, and our method outperforms in most scenarios. Furthermore, we also compare our fusion method with the only vision-based method in the physical experiment, and the results indicate that our end-to-end method empowers the robot with a more fine-grained grasp ability, reducing force redundancy by 41%. Our project is available at https://sites.google.com/view/tvgd

1. INTRODUCTION

Computer vision has become the most popular technique widely applied in perception and control problems [1]. The vision-based robotic grasp nowadays is required to fulfill different dexterous and fine-grained operations [2]. However, computer vision alone is inadequate to complete all the dexterous operations required by the grasp, especially for force-sensitive tasks [3], which inspires the idea that the tactile modality provides an emerging perceptual dimension to facilitate the robotic grasp task. Based on this, we leverage tactile and visual information to learn a tactile–visual fusion model for the fine-grained robotic grasp detection task.

Robotic grasp detection employs multiple perceptions to grasp a specific object. Conventionally, vision-based models have progressed substantially with the abundance of visual data and emerging machine-learning tools. For example, [4,5] propose typical grasp detection datasets, which are widely used in vision-based robotic grasping tasks. Some other works [6,7] adopt a vision-based dataset to predict grasp points as a regression problem. But for the limitation of vision-based methods on force-sensitive tasks [3], tactile perception becomes an emerging modality for robotic grasp detection as a supplement to vision-based methods, however, previous studies have not given a general tactile-force dataset for this task. Previous works [8–10] propose a series of Gelsight-style tactile sensors that are optic-based and superior in accuracy and texture feature extraction, but their manufacturing is complex and expensive. Other works [11,12] use electromechanical resistance based tactile sensors to obtain force information. Nevertheless, their sensors are designed for a specific task lacking versatility.

To overcome the aforementioned limitations, we propose a tactile–visual fusion based robotic grasp detection method (TVGD). Our primary contributions can be summarized as follows:

- We introduce a low-cost reproducible tactile fingertip, which can be used to sample tactile information of the fingertip conveniently and economically.
- A general tactile–visual grasp dataset is proposed, in which the basic Cornell grasp dataset [4] is extended by labeling force values on each grasp bounding box including 5,110 grasping trails.
- We propose an encoder–decoder neural network to predict affordance map for grasping including pose and force by fusing RGB and depth features hierarchically.

*Corresponding author. Email: hzsongyaoxian@163.com

- We evaluate our tactile–visual fusion method on both the public benchmark dataset and our proposed eight-object test set with different materials. Our method outperforms the benchmark and results of physical experiment show that in comparison to the only vision-based method, our fusion method can predict a fine-grained grasp reducing 41% redundant force.

2. RELATED WORK

In robotic grasp detection, conventional vision-based methods design Fully Convolutional Networks (FCNs) and Convolutional Neural Networks (CNNs) to solve grasp detection problem by supervised learning [4,6,13,14]. For tactile-based methods, as an alternative, Roberto *et al.* [15] is the first to present an end-to-end method that combines rich visual and tactile sensing, which validates the benefits of touch sensing for grasp performance. Roberto *et al.* [16] presents an end-to-end approach to learn greedy regrasping policies from raw visual–tactile data. Stephen *et al.* [17] proposes a deep tactile model predictive control (MPC), a framework for learning to perform tactile servoing from raw tactile sensor input, without manual supervision. All of them, as well as [8,18], use the Gelsight-style sensor which is an optic-based sensor. It has a high resolution, but is unable to measure force vectors directly.

Apart from the above optic-based sensors, resistance-based sensors are the other major prototype of tactile sensors. For this prototype, Sundaram *et al.* [12] designs a scalable tactile glove (STAG) to realize object identification, weight prediction, and hand pose identification by electromechanical resistance. Fang *et al.* [11] elaborates a high-density 5×5 tactile sensor array equipped on the fingertip so that the value of resistance between two electrodes will change with the ambient pressure.

Nevertheless, all the sensors aforementioned above are either custom-made and expensive, or the manufacturing process is complex and costly to fabricate. Besides, to our knowledge, there is not a general tactile dataset in public for a reproducible sensor. Comparing to existing works, our proposal in this paper has wider applicability and higher integrity, consisting of a low-cost and reproducible resistance-based sensor, a general tactile–visual dataset, and a learning-based model. Our proposed dataset is also compatible with public datasets, which can be applied in existing learning models [6,7].

3. METHOD

3.1 Grasp Definition

Grasp space representation: Conventional methods [6,7] define the grasping representation including the pose of object $p = (x, y, z, \gamma_x, \beta_y, \alpha_z)$, gripper's orientation angle ϕ, and opening width ω in Cartesian space (world/robot coordinates). For the planar grasping problem, we usually let the camera keep vertical to the tabletop, so the attitude of grasping $(\gamma_x, \beta_y, \alpha_z)$ is fixed by $(-90°, 0, 0)$ in our robot system. The final grasping representation can be defined as follows:

$$g = (x, y, z, \phi, w). \tag{1}$$

Transformation: The predicted grasp representation is usually taken place in image coordinates (pixels). We need to transform it from image coordinates to world/robot coordinates, which is split into two stages. Firstly, the transformation ${}^{Camera}_{Image}T$ from the two-dimensional (2D) image coordinates to the camera frame can be calculated by intrinsic parameters of the camera. Secondly, we obtain transformation from the camera frame to the world/robot frame ${}^{Robot}_{Camera}T$ by camera extrinsic parameters.

In image coordinates, our grasping representation can be rewritten as:

$$\tilde{g} = ((u, v), \tilde{\phi}, \tilde{\omega}), \tag{2}$$

where (u, v) is the position of the grasp center point in image coordinates. $\tilde{\phi}$ and $\tilde{\omega}$ correspond the ϕ and ω in Cartesian space (world/robot coordinates). We can obtain the grasping representation in the robot base space following the Eq. (3).

$$g = {}^{Robot}_{Camera}T \times {}^{Camera}_{Image}T \times \tilde{g}. \tag{3}$$

3.2. Problem Formulation

Conventional vision-based methods [6,7] formulate robotic grasp as a mapping modeling problem from perceptive space to grasp space:

$$G = M(I), \tag{4}$$

where $I \in \mathbb{R}^{4\times H\times W}$ denotes RGB-D images. H is the image height and W is the image width. $G = (Q, \tilde{\Phi}, \tilde{W})$ and $Q, \tilde{\Phi}, \tilde{W} \in \mathbb{R}^{H\times W}$ are each pixel's probability(**Grasp Quality**), orientation(**Angle**), and gripper's open width of the grasp in image coordinates.

However, these methods do not consider the force for grasping, which could lead to grasping failure if the force is too small, or damage the object if the force is too large. The force required for grasping an object f is described in Eq. (5), where G and μ denote the gravity and coefficient of friction for such object respectively. The coefficient of friction is related to the roughness of the surface of the object [19], which is one of the visual features. Besides, the depth image together with the RGB image could give us a hint of the size and material of the object, which implies the weight of the object. Hence, we believe the visual features could provide information on the force required for grasping an object.

$$f = \frac{G}{\mu}. \tag{5}$$

We also formulate the force prediction as a mapping modeling problem. In addition to the conventional methods, where $G = (Q, \tilde{\Phi}, \tilde{W})$, we include finger's force of the grasp $F \in \mathbb{R}^{H\times W}$, thereby leading to $G = (Q, \tilde{\Phi}, \tilde{W}, F)$.

The grasping representation is redefined as:

$$g = (x, y, z, \phi, w, f), \tag{6}$$

where (x, y, z), ϕ, w, and f are the position of the grasp, the orientation of the gripper, open width of the gripper, and the grasping force on the fingertip respectively. The (x, y, z, ϕ, w) can be obtained

by Eqs. (2) and (3), and f is predicted by M directly. The final grasp $\tilde{g}^*$ can be obtained from $\tilde{g}^* = \max_Q G$ in pixel wise.

To obtain the mapping function M, we formulate it as a regression problem. Our goal is to find a robust function M_θ to fit M:

$$\theta = \arg\min_\theta \mathcal{L}\left(G, M_\theta(I)\right), \tag{7}$$

where $\mathcal{L}$ is the loss function between the ground truth and M_θ, θ is the parameter of function M.

3.3. Network Design

To model the mapping function M, we propose a hierarchical encoder–decoder neural network M_θ to approximate it. The structure of the proposed neural network is shown in Figure 1.

We name our proposed network as **U-Grasping Network (UG-Net)** and organize it into three modules. (1) **Feature Extraction (FE):** It is in the form of a U-net [20], in which we drop the last layers of U-net and reduce the channels of each layer to one-quarter of the original numbers except the input layer. We adopt two individual branches to extract features for RGB and depth images. The features are concatenated in the decoder part. (2) **Channel-Level Attention Module (CAM)**: It is proposed in SENet [21], and we adopt to fuse two-modal features which are concatenated in **FE Module**. The module obtains $1 \times 1 \times C$ features through global average pooling (GAP), and then two fully connection (FC) layers and corresponding activation function *Relu* to build the correlation between channels, and finally outputs the weight scores of the features channel through sigmoid function. We adopt this module to fuse two-modal features efficiently compared to conventional naive 1×1 convolution operation. In our module, we also add 1×1 convolution operation to reduce the number of channels by half at last. (3) **Grasp Prediction (GP)**: It contains five grasp prediction blocks to predict grasping representation $(F, Q, (\cos(\tilde{\Phi}), \sin(\tilde{\Phi})), \tilde{W})$ respectively. Each grasp prediction block consists of two convolutional layers (Conv 3×3, *Relu*) and one linear output layer (Conv 2×2, *Linear*).

3.4 Loss Function

For our regression problem, we define our loss function as:

$$\mathcal{L} = \sum_{X \in \{Q, \tilde{\phi}, \tilde{W}, F\}} Smooth_{L1}\left(X_\theta - X^*\right), \tag{8}$$

where $Smooth_{L1}$ is formulated as:

$$Smooth_{L1}(X) = \begin{cases} 0.5(\sigma X)^2, & if\,|X| < 1 \\ |X| - 0.5/\sigma^2, & otherwise, \end{cases} \tag{9}$$

where σ is the hyperparameter that controls the smooth area, and is set to 1.0 in our work. We train the gripper orientation using an angle vector on an unit circle. Then, we rewrite the vector $(\cos(2\tilde{\Phi}), \sin(2\tilde{\Phi}))$ and Φ can be computed following Eq. (10). F^*, Q^*, $\cos(2\tilde{\Phi})^*$, $\sin(2\tilde{\Phi})^*$ and $\tilde{W}^*$ are corresponding to the ground truth.

$$\tilde{\Phi}_\theta = \frac{1}{2} arctan \frac{sin(2\tilde{\Phi}_\theta)}{cos(2\tilde{\Phi}_\theta)}. \tag{10}$$

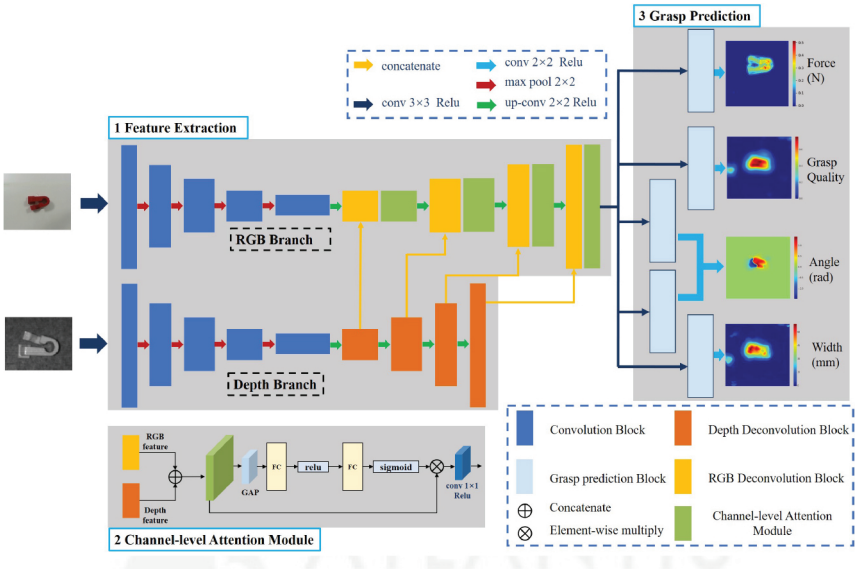


Figure 1 The structure of the neural network. The framework consists of three modules. (1) **Feature Extraction** is used to extract two-modal features. (2) **Channel-level Attention module** is to fuse two-modal efficiently. (3) **Grasp Prediction** is to predict grasping representation and five grasping prediction modules $(F, Q, (\cos(\tilde{\Phi}), \sin(\tilde{\Phi})), \tilde{W})$.

4. DATASET COLLECTION

In our work, we introduce a low-cost and reproducible tactile sensor scheme and we use this to collect a tactile–visual grasp detection dataset.

4.1 Tactile Sensor Design

We introduce a low-cost reproducible tactile fingertip shown in Figure 2 (left). The sensor consists of four contributing parts: 1. a front mount; 2. a back mount; 3. a force-sensitive film resistance; and 4. XH2.54 2pin terminal connector wire cable. All the aforementioned parts are cheap and easy to obtain. The performance of the force-sensitive film resistance is shown in Table 1. Considering the property of the force-sensitive film resistance shown in Figure 3, the digital value is linear with respect to the force value approximately following Eq. (11). We use a fundamental voltage division circuit to convert the force signal into a voltage signal and the signal is sampled by the analog-to-digital (ADC).

$$\begin{cases} U = V_{cc} \times \frac{R_s}{R_s+R_2} \sim V_{cc} \times \left(1 - k_1 \times \frac{1}{R_s}\right), \\ F = k_2 \times \frac{1}{R_s} + b, \end{cases} \tag{11}$$

where U and F are the voltage and force respectively. R_s and $R_2 = 30K\Omega$ are the force-sensitive film resistance and matching resistance respectively. V_{cc} is the 5V voltage. k_1 and k_2 are the weight coefficients. b is the bias. So we can obtain the relationship between U and F following Eq. (12). k^* and b^* are the weight coefficients. According to the sampling data, we calibrate the sensor and obtain $k^* = -0.126$ and $b^* = 5.084$ (Force unit is N and voltage unit is V) in our dataset.

$$U = k^* \times F + b^*. \tag{12}$$

4.2 Tactile–Visual Grasp Dataset

We extend the Cornell grasp detection dataset [4] with tactile information shown in Figure 4. The original dataset is a human-labeled dataset containing 885 RGB-D images of 280 different objects with ground truth labels of positive graspable rectangles and negative nongraspable rectangles. We use the proposed fingertips to label each positive graspable rectangle with a pair of force values for the specific grasping place including 5,110 grasping trails, shown in Figure 2 (Right).

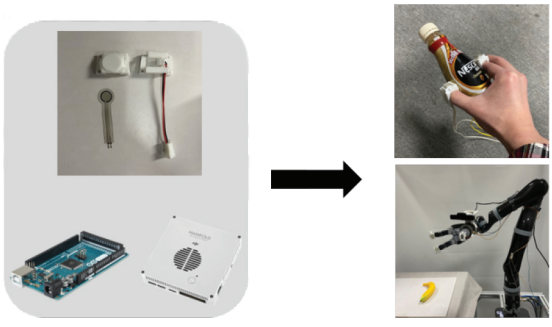


Figure 2 Our proposed low-cost reproducible tactile fingertip. (Left) The fingertip is tree- dimensional (3D)-printed and samples signal by an Arduino over robot operating system (ROS) framework. (Right) The fingertip can be fixed on a hand or robotic gripper.

5. EXPERIMENT AND EVALUATION

5.1 Data Preprocessing

We augment the dataset like most supervised methods by rotating and scaling the raw data. We scale the RGB and depth image values, and gripper's opening width values in $[0, 1]$ by normalization consistent with [6,7]. Finally, RGB and depth images are resized into 336×336 and fed into our network.

For the orientation prediction, we choose a gripper orientation angle $\tilde{\phi}$ in the range of $[-\frac{\pi}{2}, \frac{\pi}{2}]$ and represent $\tilde{\phi}$ as a vector $(\cos(2\tilde{\phi}), \sin(2\tilde{\phi}))$ on a unit circle, of which value is a continuous distribution in $[-1, +1]$. Hara *et al.* [22] shows this processing is easy for the training.

Table 1 Performance of the film resistance. The properties are satisfied with the requirement of the tactile fingertip.

Performance Index	Parameter
External diameter	16 mm
Internal diameter	10 mm
Thickness	0.24 mm
Range	0-10 kg

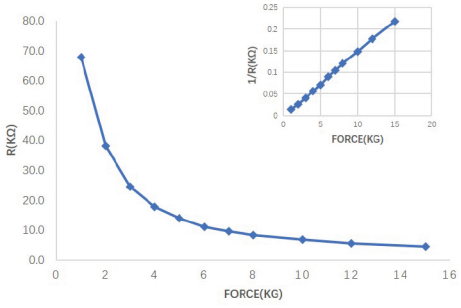


Figure 3 The characteristic curve of the force-sensitive resistance. The right-top mini-figure shows the force is linear with the reciprocal of resistance.

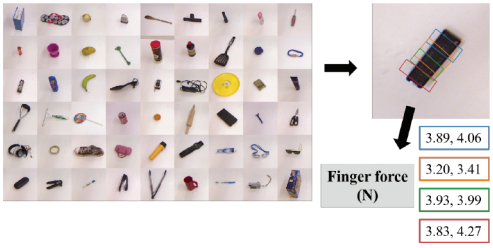


Figure 4 Tactile–visual dataset. We extend the typical Cornell grasp dataset by labeling the force value sampled from our proposed tactile fingertip. The force-value pair represents each finger force value of the typical robotic parallel jaw.

For the force prediction, considering the value distribution of the practical experiment, we find that the sensor values range from 2.5V to 5V. We scale the force value in [0, 1] by Min-Max normalization following Eq. (13). The ground truth map is made in the same way with vision-based method [6]. It is noted that since the voltage value is linear with force value based on Eq. (12), we use the voltage value to train our model.

$$x_{normal} = \frac{x - 2.5}{2.5}. \tag{13}$$

5.2 Evaluation Metrics

We introduce four metrics to evaluate the performance of our model and dataset:

Accuracy: To evaluate the accuracy of grasp prediction, we take the intersection-over-union (IoU) metric, which is widely used in previous works [6,7,23,24]. It considers a good grasp if the difference between the predicted grasp angle and ground truth angle is less than 30° and the IoU of the predicted grasp rectangle and ground truth grasp rectangle is more than 25%. The IoU (also known as Jaccard similarity coefficient) is defined following Eq. (14). Additionally, we consider a good force-based grasp if the difference between the predicted force and ground true force less than 2 N.

$$J\left(\theta, \hat{\theta}\right) = \frac{|\theta \cap \hat{\theta}|}{|\theta \cup \hat{\theta}|}, \tag{14}$$

where $\hat{\theta}$ is the predicted grasp and θ is the ground truth grasp.

Planning Time (PT): The time consumed between receiving the raw data and grasping policy generation from our network framework.

Force Quality (FQ): It measures the average force applied by the two grasping modes (with or w/o force) compared with the predicted force value. The definition is as follows:

$$FQ = \frac{\bar{F}}{F_{predict}}, \tag{15}$$

where $F_{predict}$ is the predicted force acting on the object for a single finger. $\bar{F}$ is the mean value of real-time force value sequence in **Grasp no force control** or **Regrasp with force control** phase in Figure (8a). It is noted that we average the two fingers during force recording.

Force Reduce Rate (FRR): It evaluates the degree of force redundancy of vision-based method comparing to tactile–vision fusion method and is defined by Eq. (16).

$$FRR = 1 - \frac{FQ_{with\,force}}{FQ_{no\,force}}. \tag{16}$$

5.3 Training Details

Our model is implemented with Pytorch 1.0 and contains 4.4 million (M) parameters approximately. We use the Adam optimizer to optimize the network for backpropagation during the training process. The batchsize is set to 8 and the learning rate is $lr = 1e^{-4}$. We train our model for 30 epochs. All the computation runs on a personal computer (PC) running Ubuntu16.04 with one Intel Core i7-8700K CPU and one NVIDIA Geforce GTX 1080Ti GPU.

5.4 Comparison on Dataset

Since our tactile–grasp dataset is extended from Cornell grasp dataset [4], it is fair to compare our method with other public methods on our tactile–grasp dataset, which evaluate on Conell grasp dataset originally. Our tactile–grasp dataset is divided into two different ways to evaluate the performance of the model.

- **Image-wise split:** This splits all the images in the dataset into the five folds randomly. This is to test the grasping performance on objects, which have been seen before in different poses.
- **Object-wise split:** The dataset is split based on object instances. This is to test the generation ability among different kinds of objects, which have not been seen before.

We replay and train the models on the 90% of the dataset and keep 10% to validate the performance following the above two split ways.

The comparison results are shown in Table 2*. We evaluate the performance on the vision-based method, which is the same as conventional methods [6,7]. The models are trained using visual information to predict grasping points. Furthermore, we train the proposed models using both visual and tactile information jointly and the results are shown in **{visual model}-force** entries. It is noted that **{GGCNN, GR-ConvNet-RGB-D}**-force models are derived from their original models only add a branch to predict the force map individually. **UG-Net-RGB-D** is the model that we drop the force prediction block in Figure 1. From Table 2, we can see that our models outperform in most scenarios for **Accuracy**. Especially

Table 2 Accuracy of different methods on the tactile–visual grasp dataset.

Author	Algorithm	Input Size	Accuracy (%)		PT (ms)	Parameters (Approx.)
			Image-wise	Object-wise		
Moririson [6]	GG-CNN	300 × 300	67.4	69.9	15	62k
	GG-CNN-Force		71.9	62.1	15	
Kumma [7]	GR-ConvNet-RGB-D	224 × 224	**95.5**	94.7	19	1.9 million
	GR-ConvNet-RGB-D-Force		78.7	62.0	19	
Ours	UG-Net-RGB-D	336 × 336	94.4	**96.8**	34	4.4 million
	UG-Net-RGB-D-Force		**82.0**	**75.3**	34	

*Both comparison works are replayed using the open-source code from authors' projects.

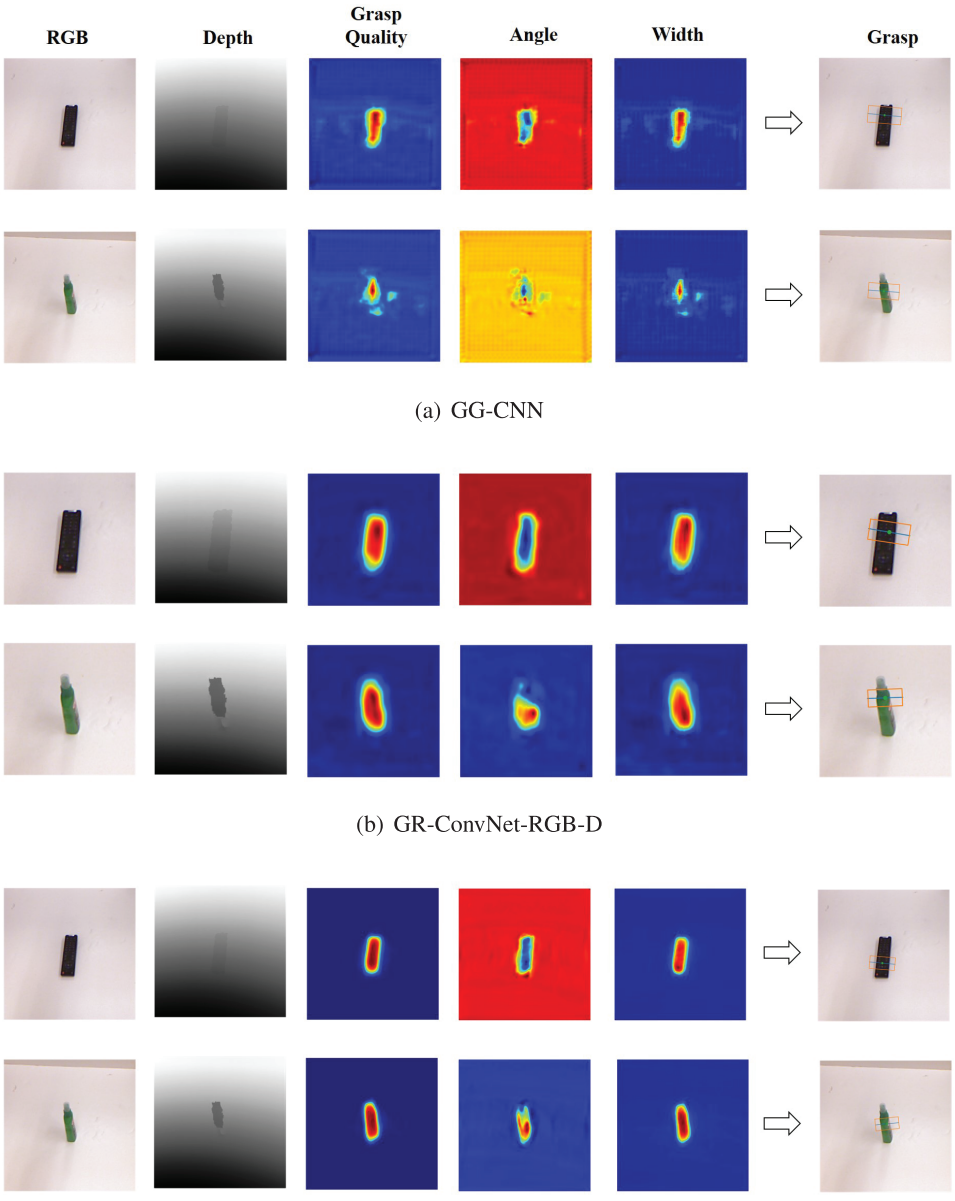
in force-predicted models, our models achieve over 4% and 13% improvement compared to other models. For **speed**, our models are slower than other models, there exist two factors. The first one is that the size of our input images is larger than others and the second one is that our model contains more parameters to train than others.

We present the visualization of both vision-based models and visual–tactile (force-predicted) models in Figures 5 and 6 respectively. From the visualization, we can see that the predictions of **GG-CNN** based models exist suboptimal points and lack of robustness. For **GR-ConvNet-RGB-D** based models, we can see that the prediction of width is too large, which needs to be improved considering the practical grasping operation. In contrast, our models can meet the robustness of the prediction and practical operations at the same time.

5.5 Physical Grasping Experiment

5.5.1 Implementation detail

For the physical experiment, we train our model using all the tactile–visual grasp dataset. We propose an eight-object test set to evaluate our tactile–visual fusion model for real robotic grasping task shown in Figure 7, of which materials are different including metal, hard plastic, rubber, wool, etc. Tested objects are placed on the tabletop randomly.

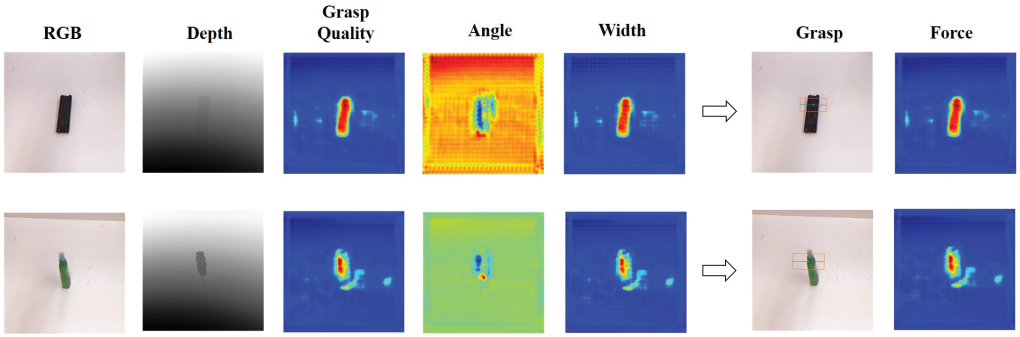


(a) GG-CNN

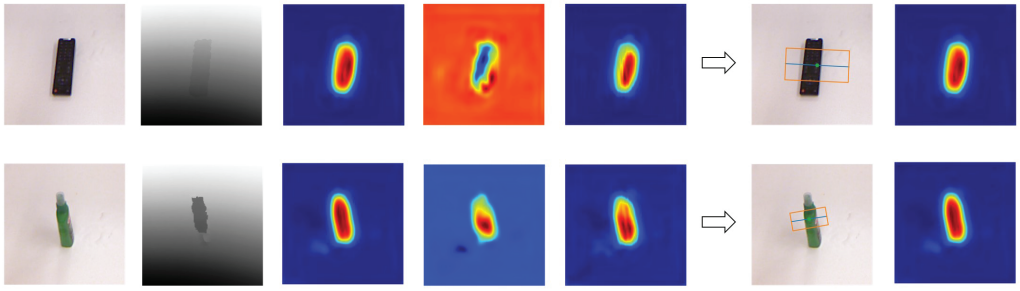
(b) GR-ConvNet-RGB-D

(c) Our method

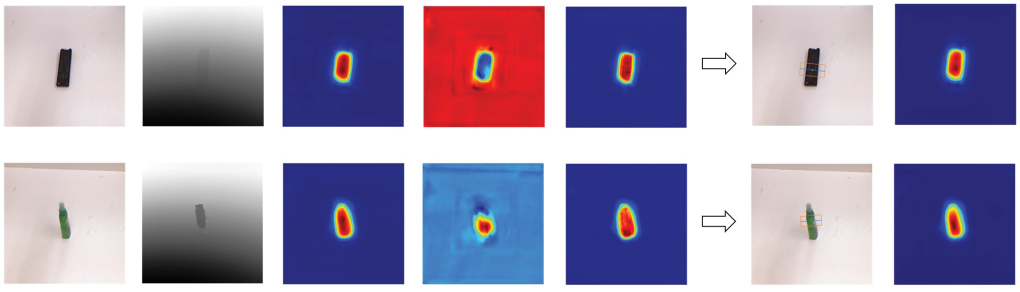
Figure 5 The visualization of vision-based models prediction. (a) The predictions from **GG-CNN** based models. (b) The predictions from **GR-ConvNet-RGB-D** based models. (c) The predictions from our proposed model.



(a) GG-CNN



(b) GR-ConvNet-RGB-D



(c) Our method

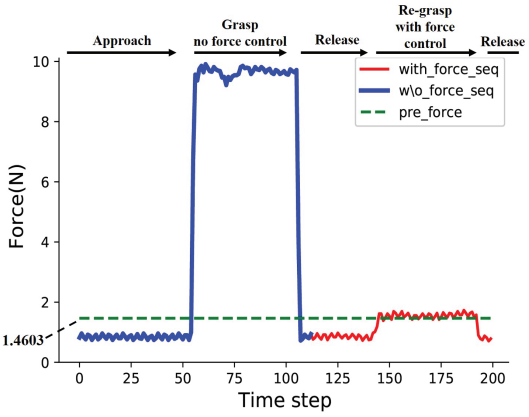
Figure 6 The visualization of vision-based models prediction. (a) The predictions from **GG-CNN** based models. (b) The predictions from **GR-ConvNet-RGB-D** based models. (c) The predictions from our proposed model.

Figure 7 Eight-object test set. The set contains different materials to evaluate our visual–tactile fusion method.

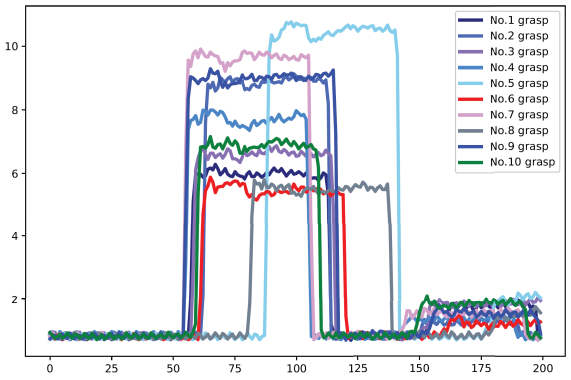
The grasping is executed by a single-arm Kinova Jaco 7DOF robot shown in Figure 2. We use an Intel RealSense SR300 RGB-D camera to obtain RGB-D images mounted on the wrist of the robot. The observation height is 55 cm away from the tabletop. We set the observation pose vertical to the tabletop approximately, which is the same as existing work [6]. Our system is running under the robot operating system (ROS) framework. We assume that the intrinsic and extrinsic parameters of the camera are known. The coordinates of RGB and depth images are aligned and the timestamps are synchronized. We obtain the raw data of RGB-D images and feed them into the model to predict an optimal grasping representation. The whole grasping pipeline based on the grasping representation is shown in Figure 8(a).

5.5.2 Grasp results

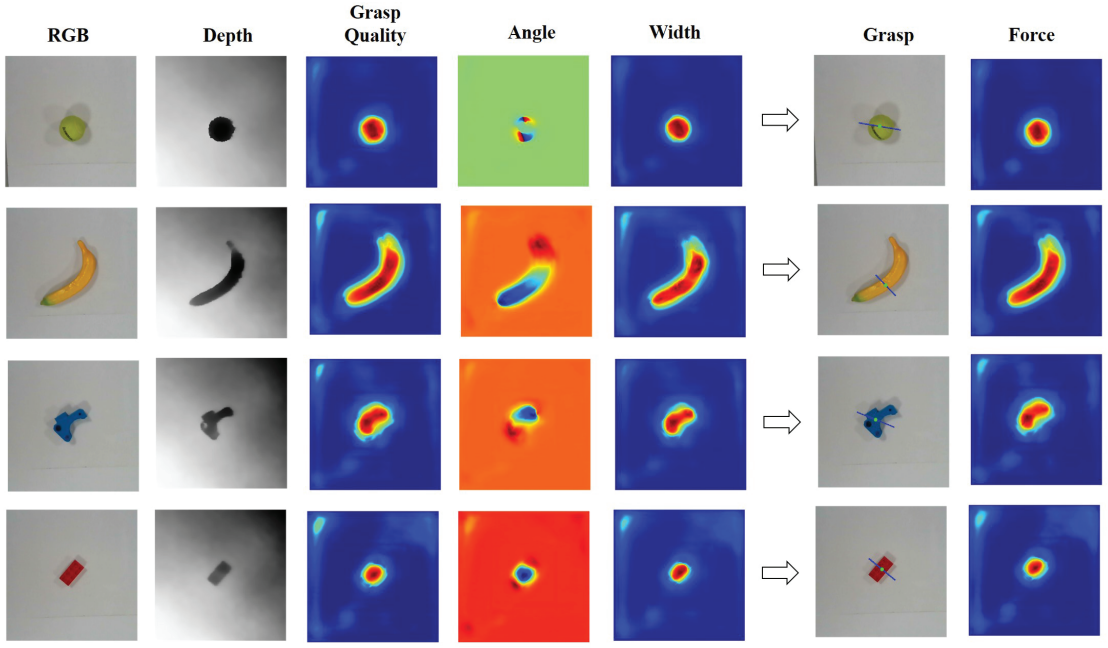
We perform vision-based and tactile–visual fusion models to grasp each object 10 times and record the pressure data from the successful grasps visualized in Figure 8(a) and (b). In the practical grasping, there are few failure cases caused by that the predicted force is too small.



(a) The whole grasping pipeline



(b) 10 grasps force sequences



(c) The prediction of our model in physical system

Figure 8 (a) The whole grasping pipeline (using a tennis ball as an example). We predict the grasping representation and grasp the target object twice. The difference between two grasps is whether the force control is banned (blue phase) or applied (red phase). The green dash line indicates the predicted finger force. (b) An example of force sequences containing 10 grasps for a tennis ball. (c) The visualization of our visual–tactile model prediction in the physical experiment.

Table 3 demonstrates the mean values of **FQ** for 10 grasping attempts from vision-based method and tactile–visual fusion method. In **w/o force** case (blue phase in Figure 8(a)), we do not set a predictive force control for the grippers, so the grippers just close directly and let the object be picked up. The force applied to the object depends on the specifications of the gripper device. In **with force** case (red phase in Figure 8(a)), the grippers grasp the object with the predictive force value generated from our tactile–visual fusion model. It can be seen that the grippers perform a smaller force to grasp the object, which proves that our model can realize a more fine-grained grasp action to avoid potential damage for force-sensitive tasks.

We observe that the value of **FQ** and **FRR** are influenced by differences in materials. The **FQ** of the vision-based method is larger than the tactile–visual fusion method obviously when the object is elastic or weak elastic. For inelastic objects, the tactile–visual fusion method also has an advantage in **FQ**. However, for materials like carton or metal, it can be seen that the values of **FQ** are all below 1.0, which possibly indicates the predicted force is too large, and there is still room for our fusion method to improve performance. In general, our method empowers a more fine-grained grasp ability than the vision-based method and reduces 41% redundant force in **FRR** averagely with a low-cost tactile sensor.

Table 3 Comparison between vision-based and tactile–visual fusion grasping results.

Object	**Tennis Ball**	**Brain**	**Banana**	**Bottle**
Property	Wool elastic	Rubber weak elastic	Bubble weak elastic	Plastic wesk elastic
FQ (w/o force)	5.048	1.288	2.210	1.322
FQ (with force)	1.076	0.891	1.229	0.921
FRR	**0.787**	0.308	**0.444**	0.303
Object	**Tetra Pak**	**T metal**	**3D printed**	**Lego**
Property	Carton inelastic	Metal inelastic	Plastic inelastic	Plastic inelastic
FQ (w/o force)	0.990	0.754	1.864	2.785
FQ (with force)	0.739	0.677	1.059	1.019
FRR	0.25	0.103	**0.432**	**0.634**

For the grasping pipeline, considering the center of gravity and materials, we keep the two grasping operations to grasp the same part of the object, which makes the results more convincing. To realize it, we grasp the object based on the vision-based model, lift and put it down to release the object vertically firstly. Then we grasp the object based on the tactile–visual fusion model again.

We present the visualization of our tactile–visual fusion model (with force) testing on the test set shown in Figure 8(c). As we can see, our model can predict the grasping point and force at the same time. It is noted that there exist some suboptimal regions in the heatmaps, which are caused by the noise of **RGB-D** camera and our background plate is not flat (which leads to the uncertainty of infrared reflection).

6 CONCLUSION

In this paper, we propose a tactile–visual fusion based robotic grasp detection method. To realize the haptic grasping, we introduce a low-cost reproducible tactile fingertip, which can be deployed on hand or robotic gripper, and use it to build a new tactile–visual grasp dataset including RGB-D and tactile information. On that basis, we propose a hierarchical encoder–decoder neural network to detect the grasp points and force in an end-to-end manner. Our method outperforms most benchmark scenarios both in the vision-based and tactile–visual fusion scheme. The physical experimental results show that our tactile–visual fusion model can make the grasp fine-grained with more suitable pressure performed on the object than the conventional vision-based method (reducing force redundancy by 41%), which enhances its applicability in force-sensitive tasks.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

AUTHORS' CONTRIBUTIONS

Yaoxian Song: The main contributor for this paper including problem formulation, proposed method, experiment, and writing. Yun Luo: data analysis and visualization. Changbin Yu: Supervision.

ACKNOWLEDGMENTS

This work was in part supported by the Major Project 2021SHZDZX0103, Pilot Project 19511132000 of Shanghai S&T Board, and the NSFC-DFG Project 61761136005.

REFERENCES

[1] E. Aguirre, M. García-Silvente, Using a deep learning model on images to obtain a 2d laser people detector for a mobile robot, Int. J. Comput. Intell. Syst. 12 (2019), 476–484.

[2] C. Haubeck, W. Lamersdorf, A. Fay, A knowledge carrying service-component architecture for smart cyber physical systems, in International Conference on Service-Oriented Computing, Springer, Málaga, Spain, 2017, pp. 270–282.

[3] J. Lee, Y.-S. Seo, C. Park, J.-S. Koh, U. Kim, J. Park, H. Rodrigue, B. Kim, S.-H. Song, Shape-adaptive universal soft parallel gripper for delicate grasping using a stiffness-variable composite structure, IEEE Trans. Ind. Electron. 2020.

[4] I. Lenz,H. Lee, A. Saxena, Deep learning for detecting robotic grasps, Int. J. Robot. Res. 34 (2015), 705–724.

[5] A. Depierre, E. Dellandréa, L. Chen, Jacquard: a large scale dataset for robotic grasp detection, in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Madrid, Spain, 2018, pp. 3511–3516.

[6] D. Morrison, P. Corke, J. Leitner, Learning robust, real-time, reactive robotic grasping, Int. J. Robot. Res. 39 (2020), pp. 183–201.

[7] S. Kumra, S. Joshi, F. Sahin. Antipodal robotic grasping using generative residual convolutional neural network, in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Las Vegas, USA, 2020, pp. 9626–9633.

[8] R. Li, E.H. Adelson, Sensing and recognizing surface textures using a gelsight sensor, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013, pp. 1241–1247.

[9] W. Yuan, S. Dong,E.H. Adelson, Gelsight: high-resolution robot tactile sensors for estimating geometry and force, Sens. 17 (2017), p. 2762.

[10] B. Fang, F. Sun, C. Yang, H. Xue, W. Chen, C. Zhang, D. Guo, H. Liu, A dual-modal vision-based tactile sensor for robotic hand grasping, in 2018 IEEE International Conference on Robotics

and Automation (ICRA), IEEE, Brisbane, Australia, 2018, pp. 4740–4745.
[11] B. Fang, F. Sun, Y. Chen, C. Zhu, Z. Xia, Y. Yang, A tendon-driven dexterous hand design with tactile sensor array for grasping and manipulation, in 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), IEEE, Dali, China, 2019, pp. 203–210.
[12] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, W. Matusik, Learning the signatures of the human grasp using a scalable tactile glove, Nature. 569 (2019), pp. 698–702.
[13] R. Detry, C.H. Ek, M. Madry, D. Kragic, Learning a dictionary of prototypical grasp-predicting parts from grasping experience, in 2013 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Karlsruhe, Germany, 2013, pp. 601–608.
[14] D. Kappler, J. Bohg, S. Schaal, Leveraging big data for grasp planning, in 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Seattle, USA, 2015, pp. 4304–4311.
[15] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E.H. Adelson, S. Levine, The feeling of success: does touch sensing help predict grasp outcomes?, in Proceedings of (CoRL) Conference on Robot Learning, 2017, pp. 314–323. http://proceedings.mlr.press/v78/calandra17a.html
[16] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, S. Levine, More than a feeling: learning to grasp and regrasp using vision and touch, IEEE Robot. Autom. Lett. 3 (2018), pp. 3300–3307.
[17] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, S. Levine, Manipulation by feel: touch-based control with deep predictive models, in 2019 International Conference on Robotics and Automation (ICRA), IEEE, Montreal, Canada, 2019, pp. 818–824.
[18] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, S. Levine, Omnitact: a multi-directional high-resolution touch sensor, in 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Virtual, 2020, pp. 618–624.
[19] J. Bikerman, Surface roughness and sliding friction, Rev. Mod. Phys. 16 (1944), 53.
[20] O. Ronneberger, P. Fischer, T. Brox, U-net: convolutional networks for biomedical image segmentation, in International Conference on Medical Image Computing and Computer-Assisted Intervention,Springer, Munich, Germany, 2015, pp. 234–241.
[21] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018, pp. 7132–7141.
[22] K. Hara, R. Vemulapalli, R. Chellappa, Designing deep convolutional neural networks for continuous object orientation estimation, *arXiv preprint arXiv:1702.01499,* 2017.
[23] J. Redmon, A. Angelova, Real-time grasp detection using convolutional neural networks, in 2015 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Seattle, USA, 2015, pp. 1316–1322.
[24] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, N. Zheng, Roi-based robotic grasp detection for object overlapping scenes, in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, Macao, China, 2019, pp. 4768–4775.