2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Grasp State Assessment of Deformable Objects Using Visual-Tactile Fusion Perception

Shaowei Cui[1,2], Rui Wang[1], Junhang Wei[1,2], Fanrong Li [1,2], and Shuo Wang[1,3,*]

Abstract— **Humans can quickly determine the force required to grasp a deformable object to prevent its sliding or excessive deformation through vision and touch, which is still a challenging task for robots. To address this issue, we propose a novel 3D convolution-based visual-tactile fusion deep neural network (C3D-VTFN) to evaluate the grasp state of various deformable objects in this paper. Specifically, we divide the grasp states of deformable objects into three categories of sliding, appropriate and excessive. Also, a dataset for training and testing the proposed network is built by extensive grasping and lifting experiments with different widths and forces on 16 various deformable objects with a robotic arm equipped with a wrist camera and a tactile sensor. As a result, a classification accuracy as high as** 99.97% **is achieved. Furthermore, some delicate grasp experiments based on the proposed network are implemented in this paper. The experimental results demonstrate that the C3D-VTFN is accurate and efficient enough for grasp state assessment, which can be widely applied to automatic force control, adaptive grasping, and other visual-tactile spatiotemporal sequence learning problems.**

I. Introduction

Robotic grasp capability is receiving increasing attention due to increased demand for various dexterity grasping and manipulation of service and industrial robots [1], [2]. To improve the general grasp ability of the robots, accurate and efficient grasp state assessment is a relatively critical part. Traditional grasp quality assessment focuses on whether a grasp process is stable and whether slippage has occurred. Many scholars have already researched in the grasp stability assessment [3], [4] and slip detection/prediction [5], [6].

Nevertheless, for a deformable or fragile object, it is not enough to only detect whether it slides during a grasp process. For example, for such a task of grasping paper cups, if the gripping force is set too large, although the paper cup can be prevented from slipping, the excessive gripping force may cause the paper cup to undergo a large deformation, thereby causing irreversible damage. To this end, a more comprehensive approach to assess the grasp state of deformable objects needs to be studied. In this paper, we define the grasp state assessment task for deformable objects as a tri-classification problem with sliding, appropriate, and excessive labels. These three grasp states are used to describe the grasp state of various deformable objects, as shown in Fig. 1.

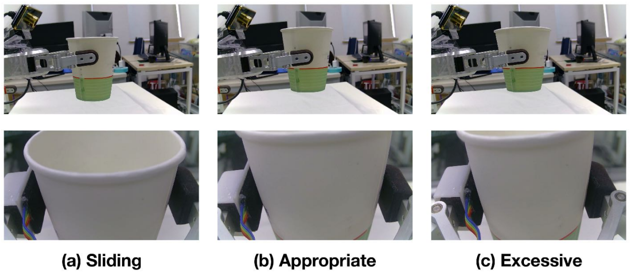


Fig. 1: (a): The sliding grasp state. (b) The appropriate grasp state. (c) The excessive grasp state. The top row images are captured by a side-mounted camera and the bottom by a wrist camera.

Vision and tactile sensing are two of the primary sensing modalities to perceive the ambient world for humans [7]. Vision provides the appearance, shape, and other visible features of objects, while touch provides more accurate texture, roughness, contact strength, and other invisible details [8]. For such a grasp state assessment task, humans are capable of intuitively performing the evaluation process. Someone picking up a random object can automatically determine if the grasp is appropriate [9]. This information benefits from both tactile and visual feedback. The same is true for robots, and this paper focuses on how to endow robots the ability to evaluate the grasp state of deformable objects using visual-tactile fusion perception.

The difficult primary issue involved in such a bimodal fusion perception task is how to learn effective fusion spatiotemporal features from two heterogeneous modal spatiotemporal sequences [5]. In this paper, we propose a novel 3D convolution-based visual-tactile fusion deep neural network (C3D-VTFN) to evaluate the grasp state, mimicking the strategy adopted by humans. Furthermore, we perform extensive grasping and lifting experiments with different grasp settings to train and test the neural work on our humanoid robot platform. The visual and tactile sequences are taken from a wrist camera fixed above a gripper and a XELA [10] tactile sensor, respectively. The experimental setup is shown in Fig. 2. Finally, some comparative experiments of C3D-VTFN model with different inputs and two real-time grasp state correction experiments based on proposed model are implemented to verify the effectiveness of the proposed network further.

This paper is organized as follows: in Section II, the

This work was supported in part by the National Natural Science Foundation of China under Grant U1713222, Grant 61773378. (corresponding author: Shuo Wang shuo.wang@ia.ac.cn)

[1] S. Cui, R. Wang, J. Wei, F. Li, and S. Wang are with Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

[2] S. Cui, J. Wei, and F. Li are with School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China.

[3] S. Wang is also with Center for Excellence in Brain Science and Intelligence Technology Chinese Academy of Sciences, Shanghai 200031, China.

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

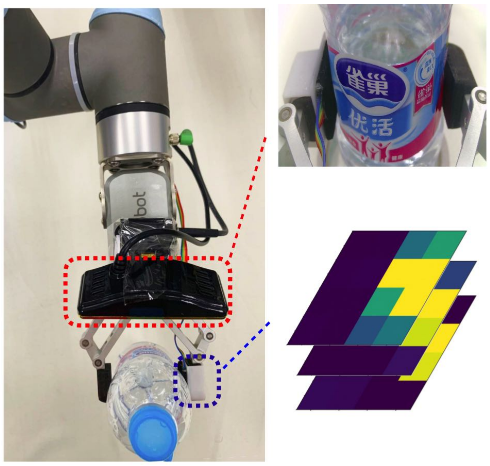


Fig. 2: Left: The experiment setup: a UR3 robot arm with an OnRobot RG2 gripper. One finger of the gripper is equipped with a XELA Tactile sensor [10]. A 1080P USB camera is mounted on the top of the gripper. Upper right: The photo taken by the wrist camera. Bottom right: Three-axis force distribution map from the tactile sensor.

related work of grasp state assessment and visual-tactile fusion perception are explained. In Section III, the problem statement, detailed architecture of C3D-VTFN, and training specifications are described. In Section IV, the experimental results and discussions are provided. Finally, the contributions of this paper and future work are discussed in Section V.

II. Related work

A. Grasp state assessment

Grasp state assessment is critical for robots to achieve high-quality grasping and manipulation tasks. In the past decades, most studies have focused on the stability in the grasp process. Yasemin Bekiroglu *et al.* [11] studied the problem of learning grasp stability in robotic object grasping based on tactile measurement and Hidden Markov Models (HMMs). [12] proposed an integrating grasp planning with an online stability assessment based on tactile sensing. They also presented a probabilistic framework for grasp modeling and stability assessment [13]. Yevgeb Chebotar *et al.* introduced a framework for learning re-grasping behaviors based on tactile data. They presented a grasp stability predictor that used spatio-temporal tactile features [14].

Moreover, a novel method to incorporate exteroception and proprioception into grasp stability assessment was proposed by [3]. A convolutional neural network (CNN) was used to extract features and fusion of different modality information. More recently, A new method to predict grasp stability using a non-matrix tactile sensor was proposed by [15]. Filipe Veiga *et al.* proposed a grip stabilization approach for novel objects based on slip prediction [6]. Besides, Graph convolutional network method was also studied to predict grasp stability with tactile sensors[16].

Most of the above studies have focused on stability assessment during the grasp process while ignoring dexterity. However, excessive-force gripping can achieve stable grasp of deformable objects but may also cause irreversible damage. Therefore, we conduct a more comprehensive grasp state assessment of daily life deformable objects. By adding the grasp state of excessive detection, a more comprehensive grasp state assessment framework is proposed for more sophisticated and complex robotic grasping and manipulation tasks.

B. Visual-tactile fusion perception

Vision and tactile sensing are two primary important modalities in robotics perception. In the past decades, it is still challenging to combine vision and touch modalities to facilitate robot manipulations due to their different sensing principles and data structures. However, these limitations have recently improved due to the increasing measurement accuracy of tactile sensors and advances in fusion algorithms with deep neural networks. The combination of visual and tactile perception plays an increasingly important role in the robotic community [17]. Visual-tactile fusion perception has long been used for a variety of tasks, such as surface classification [18], object recognition [19], object 3D shape perception [20], etc.

In the field of grasping and manipulation, R Calandra *et al.* investigated the question of whether touch sensing aids in predicting grasp outcomes within a multimodal sensing framework that combines vision and touch [17]. The experimental results indicated that incorporating tactile readings substantially improve grasp performance. Furthermore, an end-to-end action-conditional model that learns re-grasping policies from rowed visual-tactile data was proposed in [21]. The re-grasping strategy using combined visual and tactile sensing had greatly improved the success of grasping. Michelle A. Lee *et al.* used self-supervision to learn a compact and multimodal representation of RGBs, depth, force-torque, and proprioception for different contact-rich manipulation [22].

Nevertheless, these studies only use tactile and visual images at a specific moment as input and do not use time-domain information of the two modalities. Spatiotemporal features of visual and tactile are extracted by Convolutional Neural Network (CNN)+Recurrent Neural Network (RNN) architecture for slip detection [5]. However, they only detect slip, and the premise of this study is that the reading frequency of the tactile and visual data is consistent, but most tactile sensors read more quickly than the cameras. Hence, we present a novel C3D grounded framework to tackle the visual-tactile fusion perception problem.

III. Proposed method

A. Problem statement

Our goal is to obtain the current grasp state by visual-tactile fusion perception. Given the visual $(X_{v1}, X_{v2},...,X_{vm})$ and tactile $(X_{t1}, X_{t2},...,X_{tn})$[1] sequences, we first extract visual

[1] m and n are the lengths of the two sequences, respectively.

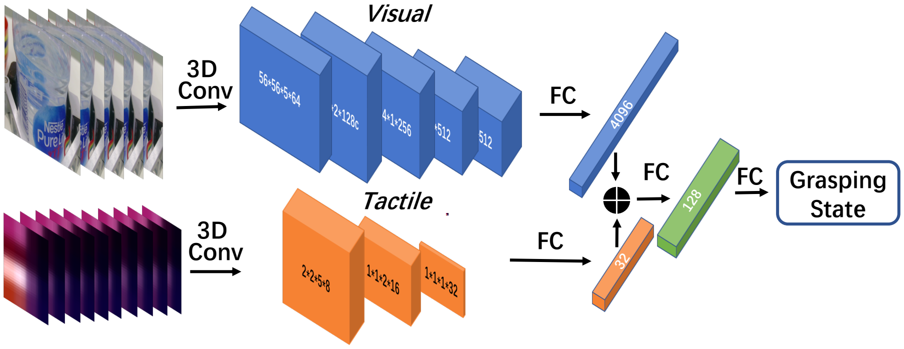


Fig. 3: The diagram of C3D-VTFN model. Blue and orange blocks denote visual and tactile 3D convolutional layers, respectively. The cuboids with different colors represent FC layers.

(F_v) and tactile (F_t) features by visual (E_v) and tactile (E_t) encoder functions and then construct a fusion feature ($F_{v,t}$) based on them. Finally, $F_{v,t}$ is fed into a classification function $\mathbb{F}_c$ to predict the current grasp state y. This problem is formulated as

$$F_{v,t} = E_v(X_{v1}, X_{v2}, .., X_{vm}) \oplus E_t(X_{t1}, X_{t2}, .., X_{tn}) \tag{1}$$

$$y = \mathbb{F}_c(F_{v,t}) \;\; y \in 0, 1 2 \tag{2}$$

Which 0, 1, and 2 refer to the sliding, appropriate, and excessive grasp states, respectively. Hence, the grasp state assessment task is defined as a tri-classification problem.

To address the above problem, we propose a novel 3D Convolution-based visual-tactile fusion network (C3D-VTFN) in this paper, where E_v and E_t are implemented by 3D convolutional neural networks with parameters θ_v and θ_t, and $\mathbb{F}_c$ is constructed by Fully-Connection (FC) layers with parameters θ_c.

B. Model description

The overall architecture of C3D-VTFN model is shown in Fig. 3. The proposed model consists of three components including visual feature extraction module (E_v), tactile feature extraction module (E_t), and classification module ($\mathbb{F}_c$). Given the current visual and tactile sequence, the output of C3D-VTFN is the current grasp state category. Firstly, the visual and tactile features of each spatiotemporal sequence are extracted by C3D networks. Note that tactile modal input is also treated as a small image because of its matrix distribution. Finally, the visual and tactile features are then combined using FC layers to generate a classification result.

In practice, we use five [2] $112 \times 112 \times 3$ [3] visual images and ten $4 \times 4 \times 3$ tactile images as input and the detailed network parameters are shown in Table I. The visual features extraction module includes five C3D and two FC layers. The convolution kernel size and stride size of each convolutional layer are not exactly the same. The output size of the final visual C3D layer is $4 \times 4 \times 1 \times 512$. The features from the visual C3D layers are fed to two FC layers and transformed into a 4096-dimensional feature vector. Similarly, the tactile features extraction module are composed of three C3D layers and followed by two FC layers. The difference is that the output tactile feature vector is 32-dimensional. Finally, the feature vectors from the two modalities are concatenated together to output the final grasp state category through two classification FC layers.

TABLE I: Detailed network parameters of C3D-VTFN.

	Visual layers	**Output size**
3d-conv$_1$	3×3×3×64, padding(1,1,1), relu	112×112×5×64
pool$_1$	Max(1,2,2), stride (1,2,2)	56×56×5×64
3d-conv$_2$	3×3×3×128, padding(1,1,1), relu	56×56×5×128
pool$_2$	Max(2,2,2), stride (2,2,2)	28×28×2×128
3d-conv$_{3a}$	3×3×3×256, padding(1,1,1), relu	28×28×2×256
3d-conv$_{3b}$	3×3×3×256, padding(1,1,1), relu	28×28×2×256
pool$_3$	Max(2,2,2), stride (2,2,2)	14×14×1×256
3d-conv$_{4a}$	3×3×3×512, padding(1,1,1), relu	14×14×1×256
3d-conv$_{4b}$	3×3×3×512, padding(1,1,1), relu	14×14×1×256
pool$_4$	Max(1,2,2), stride (1,2,2)	7×7×1×512
3d-conv$_{5a}$	3×3×3×512, padding(1,1,1), relu	7×7×1×512
3d-conv$_{5b}$	3×3×3×512, padding(1,1,1), relu	7×7×1×512
pool$_5$	Max(1,2,2), str(1,2,2), pad(1,0,0)	4×4×1×512
fc$_1$	(8,192, 4,096)	1×1×4,096
fc$_2$	(4,096, 4,096)	1×1×4,096
	Tactile layers	**Output size**
3d-conv$_6$	3×3×3×8, padding(1,1,1), relu	4×4×10×8
pool$_6$	Max(2,2,2), stride (2,2,2)	2×2×5×8
3d-conv$_7$	3×3×3×16, padding(1,1,1), relu	2×2×5×16
pool$_7$	Max(2,2,2), stride (2,2,2)	1×1×2×16
3d-conv$_8$	3×1×1×32, padding(1,0,0), relu	1×1×3×32
pool$_8$	Max(2,1,1), stride (2,1,1)	1×1×1×32
	Classification layers	**Output size**
fc$_3$	(4096+32,128)	1×1×128
fc$_4$	(128, 3) max	1×1×1

Specifically, we use Xavier initialization [23] to initiate network weights and cross-entropy [24] as the loss function. Adam optimizer [25] with $1e-07$ learning rate is adopted in the training process. The model is implemented on the PyTorch platform [4] and trained on an NVIDIA DGX server. The batch size is set as 8 in this paper.

[2] The sequence length was selected by comparison experiments.

[3] We selected this size as the default visual modal input size by comparing the performance of the model with different image sizes as input.

[4] Source code for study replication is available at: `https://github.com/swchui/Grasping-state-assessment`

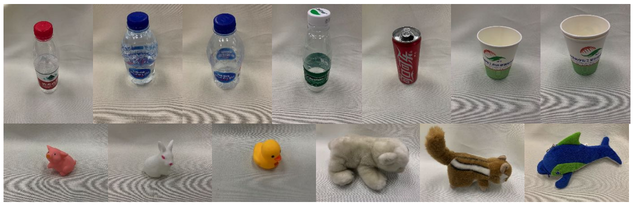


Fig. 4: Some deformable objects of the GSA dataset.

IV. Experiments

In this section, we first introduce our grasp state assessment dataset (GSA dataset) and the experimental setup. Then the performance comparison of C3D-VTFN with different structures and parameters on the GSA Dataset is provided. Finally, we perform two delicate grasp experiments of a deformable object based on the C3D-VTFN model.

A. The GSA dataset introduction

All of the experiments are conducted with a 6-DOFs UR3 robot arm equipped with an OnRobot RG2 gripper. Specifically, one finger of the gripper is covered by a XELA tactile sensor and a 1080P USB camera is mounted on the top of the gripper as a wrist-camera. The robot setup is shown in Fig. 2.

The GSA dataset is built by extensive grasping and lifting experiments on 16 deformable objects of different sizes, shapes, textures, materials, and weights, some of them are shown in Fig. 4. Inspired by [5], different grasp widths and forces are selected to balance the number of routines with different labels. In this way, the grasp states are automatically labeled in each grasping and lifting trial. In each grasp experiment, an object is grasped with the preset width and force and lifted slowly for 20.0 mm (The lifting speed is set to 10.0 mm/s). During the grasping and lifting process, the data are collected by the visual sensor with a 30 Hz and tactile sensor with 60 Hz, respectively. We perform 50 to 60 grasps per object, collecting approximately 30 to 40 frames of visual images and 60 to 80 frames of tactile images per grasp trial. As a result, the GSA dataset consists of approximately 20,000 5-frame visual image sequences and corresponding tactile image sequence samples. Among them, the grasping data of randomly selected thirteen objects is used to train the model, and the grasping data of the remaining three objects is used for testing. The detailed GSA dataset is available at `https://github.com/swchui/Grasping-state-assessment/graspingdata`.

B. Performance comparison results

To evaluate the performance of the proposed model more comprehensively and accurately, we compared the Precision, Recall, F1 score, and model size of the model with different inputs and structures. The Precision, Recall, and F1 score are used to evaluate the classification performance, and the model size is adopted to compute the real-time performance of different models. The performance of the model may be affected by different length of input sequence, image size, inter-frame interval, and input of single-mode or dual-mode perception.

1) Different input lengths: Longer sequences not only mean more temporal information, but also result in redundant calculations and bring more noise information. The visual sequence lengths of 3, 4, 5, 6, 7, and 8 are selected as the inputs of the model for comparative evaluation, and the tactile sequence length is set to twice the visual according to the reading frequency. The experiments results are shown in Table II.

TABLE II: Experimental results of the models with different input length.

	Sequence length					
	3	4	**5**	6	7	8
Precision	75.78	95.21	**99.97**	99.80	90.27	90.13
Recall	67.27	95.49	**99.98**	99.74	88.81	79.91
F1 score	67.42	95.08	**99.98**	99.77	88.79	79.83
Size (M)	78.53	78.53	78.53	78.54	78.54	78.54

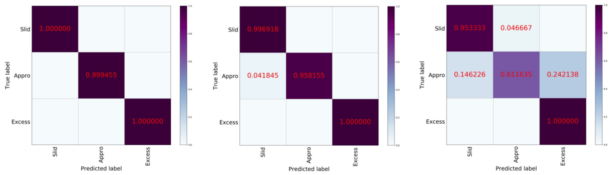
The results show that it is not the more extended the input sequence is, the better the classification performance is. The model with a sequence length of 8 has a classification accuracy of 10% lower than that of the sequence length of 5. As a result, the optimal classification performance is obtained when the sequence length is 5.

2) Different inter-frame interval: Since sequences with different time intervals have different characteristics, the large inter-frame interval can result in a reduction in sample rate, and whether this affects the performance of the model is still a question worth exploring. Due to the reading speed of sensors is fixed, we set a *basic* input sample in which the visual and tactile images are consecutive recorded reading. We also build *step* input samples in which the data reading is selected with step 2 and 3. We set the other parameters as default, and only change the inter-frame interval for a comparison test. The results are shown in Table III.

TABLE III: Experimental results of the models with different input inter-frame intervals.

	Frame interval		
	1	2	3
Precision	**99.97**	98.03	85.62
Recall	**99.98**	98.50	85.50
F1 score	**99.98**	98.23	83.80
Size (M)	78.53	78.53	78.53

Table III suggests that the using *step* sampling method would be worse, especially for the appropriate grasping state. The confusion matrix shows that the grasp state of the appropriate is higher in the case where the step setting is smaller, as shown in Fig. 5. The intuitive explanation is that the reduction of the sample rate will reduce the confidence of the proposed model in determining the proper grasp state, which makes it more biased toward sliding or excessive state.



(a) Basic (b) Step 2 (c) Step 3

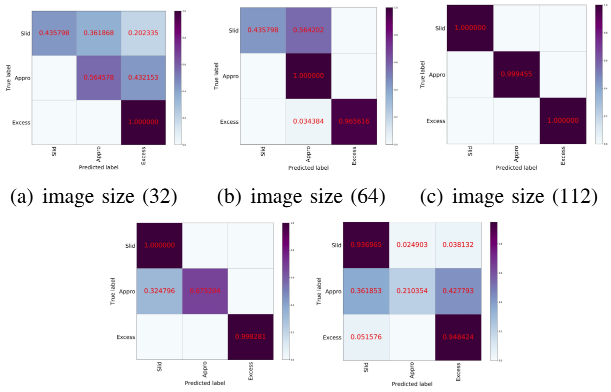
Fig. 5: The confusion matrices of models with different inter-frame interval.

3) Different image size: The image size directly determines the amount of information input by the visual modality and the amount of model parameters. Therefore, we set the inter-frame interval to 1, set the input image sequence length to 5 (the tactile sequence length corresponds to 10), set the input image size to 32×32, 64×64, 112×112, 224×224, and 512×512 respectively, and modify the corresponding model parameters, the results are shown in Table IV.

TABLE IV: Experimental results of the models with different image size.

	Image size				
	32	64	**112**	224	512
Precision	76.78	90.01	**99.97**	89.36	73.44
Recall	66.68	80.05	**99.98**	89.12	69.86
F1 score	66.48	80.44	**99.98**	87.21	62.23
Size (M)	**55.53**	64.37	78.53	92.69	106.84

The experimental results indicate that the model performance is not directly proportional to image size. We find that as the size of the image increases, the model detects the sliding state more accurately, but the appropriate grasp state detection performance becomes worse, as shown in Fig. 6. Hence, we select the visual sequence with image size 112 as the input to the model.



(a) image size (32) (b) image size (64) (c) image size (112)

(d) image size (224) (e) image size (512)

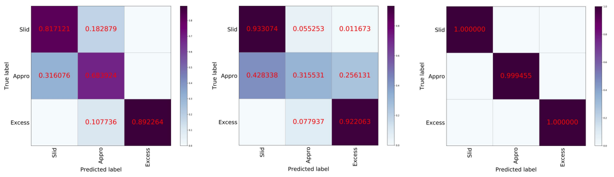
Fig. 6: The confusion matrices of models with different image size.

4) Single-or-dual modal perception: Furthermore, different modal combinations are tested to verify the performance benefits of visual-tactile fusion perception. For the visual-only mode, we select the visual features extraction module and classification module from the C3D-VTFN model (as shown in Fig. 3). Similarly, the tactile-only mode is tested by combining the tactile features extraction module and a classification module. The comparison results are shown in Table V.

TABLE V: Experimental results of the models with single-or-dual modal perception.

	Tactile-only	Visual-only	Visual-Tactile fusion
Precision	70.30	79.74	**99.97**
Recall	72.36	79.77	**99.98**
F1 score	67.11	79.27	**99.98**
Size (M)	**0.01**	78.52	78.53

According to the experimental results, visual-tactile fusion perception achieves much better precision, recall, and F1 score than that of any single modal perception. Theoretically, the visual image provides the geometrical information of the contact situation, which would have a better ability to distinguish the excessive grasp state from others. This analysis has been verified by the confusion matrices shown in Fig. 7(a) and Fig. 7(b). Meanwhile, the confusion matrices also show that the tactile-only model achieves better detection performance of the sliding grasp state than the visual-only model.



(a) Visual-only (b) Tactile-only (c) VTF

Fig. 7: The confusion matrices of models with different inputs.

C. Delicate grasp experiments based on C3D-VTFN model.

Two delicate grasp experiments in this section are performed to further verify the effectiveness of the proposed model. We have developed a roughly grasp adjustment strategy that adjusts grasp width and force in real-time based on the grasp state detector (C3D-VTFN). The detailed grasp regulation strategies are as follows,

$$w_{t+1}, f_{t+1} = \begin{cases} w_t - 1, f_t + 1 & c_t = sliding, \\ w_t, f_t & c_t = appropriate, \\ w_t + 1, f_t + 1 & c_t = excessive. \end{cases}$$

Where w_t and f_t represent the grasp width and force at the current moment, and w_{t+1} and f_{t+1} represent the next moment. Also, the adjustment of the grasp settings depends on the evaluation of the grasp state of the current moment c_t.

On the one hand, a grasp adjustment experiment that begins with a slip is first performed, as described in Section

IV-C.1. On the other hand, a grasp adjustment experiment with an initial excessive-force grip is performed, as described in Section IV-C.2.

1) The sliding grasp experiments: First, we set the grasp force to 5N and the grasp width to 66 mm for the grasping and lifting experiments of a deformable bottle (not included in the GSA dataset) with (WA) and without (WoA) adjustment strategy. In these two experiments, the real-time changing curve of the adjustment grasp force, grasp width, and the detection grasp state is shown in Fig. 8.

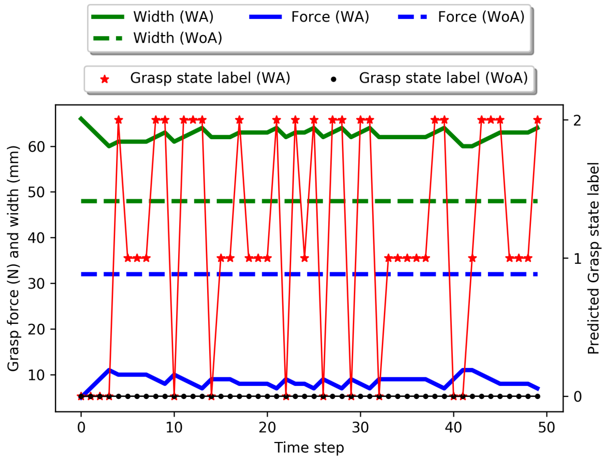


Fig. 8: The real-time changing curve of different values in the two sliding grasp experiments. Solid line: the grasp process with adjustment. Dashed line: the grasp process without adjustment. The second axis labels: (0) sliding, (1) appropriate, (2) excessive.

Fig. 8 shows that the proposed model can accurately detect the current state of the grasp state regardless of there is a grasp strategy adjustment. In the experiment without the adjustment strategy, the detector always detected the grasp state as *sliding (0)*. However, in the experiment with the adjustment strategy, the detection state changes with the change of the grasp settings. In the grasp process with adjustment, the model still detects it as a sliding state even if has been adjusted at the beginning. The reason is that the preset grip width (68mm) is larger than the actual diameter of the bottle and takes a few steps to adjust. After a few steps, the entire lifting process is adjusted between the three grasp states according to the actual grasp situation to achieve a delicate grasp.

2) The excessive grasp experiments: Similar to the previous experiments, we preset the grasp force and width to 32N and 48mm, respectively, in the excessive grasp experiments. The changing curves of different grasp values are shown in Fig. 9.

The grasp process in the experiment with adjustments can finally stabilize the grasp force and width, similar to those in the sliding experiment. Since the grasp force and width adjustment settings are very rough in this paper, the final steady-state of the two experiments is not wholly consistent. But it is sufficient to verify the performance of the proposed model. Please note that the retardation of the predicted grasp state at the beginning of the grasp adjustment experiment is due to the experimental settings and the initial grasp force setting being too large.

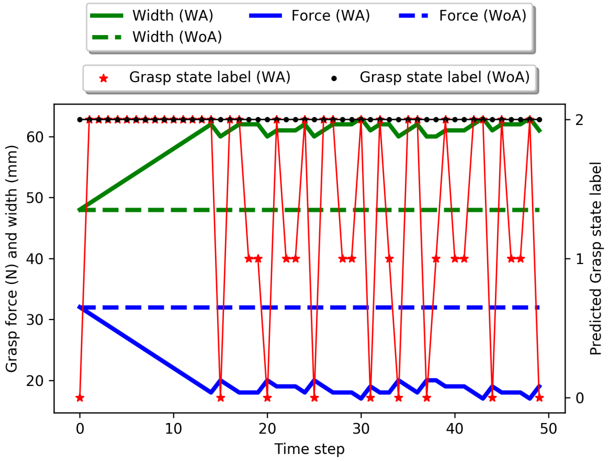


Fig. 9: The real-time changing curve of different values in the two excessive grasp experiments.

The above two experiments have well verified the evaluation performance of the proposed model on the current grasp state. However, this adjustment strategy is not enough for delicate grasp, and it is necessary to set a more fine-grained adjustment strategy (See supplementary video materials for more experimental details).

Additionally, due to the fixed angle of view during data collection, the trained model has higher accuracy at fixed angles and less than ideal performance at other views. A feasible way is adding multiple different views of grasp settings in each experiment, which significantly increases the scale of the GSA dataset, and makes the model more generalized. Fortunately, this limitation does not prevent us for verifying the feasibility of the C3D-based visual-tactile fusion perception approach.

V. Conclusion and Future Work

A network named C3D-VTFN is proposed to assess the grasp state of various deformable objects by using visual-tactile fusion perception in this paper. We extract the features of the visual and tactile modalities by 3D convolution layers, which provides a new feature extraction scheme for the visual-tactile fusion perception tasks. Besides, the GSA dataset used to train and test the proposed model is established by extensive grasping and lifting experiments in this paper, and the experimental results show the effectiveness and high accuracy of the proposed model. Finally, we perform two delicate grasp experiments with a rough adjustment strategy based on the proposed model and achieved convincing results.

In the future, we will explore perceptual models that are more in line with human vision-tactile fusion properties and their applications in robotic grasping and manipulation.

REFERENCES

[1] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D.Quillen, “Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection,” *The International Journal of Robotics Research*, vol. 37, no. 4–5, pp. 421–436, Apr. 2018.

[2] J. Sanchez, J. Corrales, B. Bouzgarrou, and Y. Mezouar, “Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey,” *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, Jun. 2018.

[3] J. Kwiatkowski, D. Cockburn, and V. Duchaine, “Grasp stability assessment through the fusion of proprioception and tactile signals using convolutional neural networks,” in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Vancouver, Canada, Sep. 2017, pp. 286–292.

[4] D. Cockbum, J. P. Roberge, A. Maslyczyk, V. Duchaine, *et al.*, “Grasp stability assessment through unsupervised feature learning of tactile images,” in *2017 IEEE International Conference on Robotics and Automation*, Singapore, Singapore, May. 2017, pp. 2238–2244.

[5] J. Li, S. Dong, and E. Adelson, “Slip detection with combined tactile and visual information,” in *2018 IEEE International Conference on Robotics and Automation*, Brisbane, Australia, May. 2018, pp. 7772–7777.

[6] F. Veiga, J. Peters, and T. Hermans, “Grip stabilization of novel objects using slip prediction,” *IEEE transactions on haptics*, vol. 11, no. 4, pp. 531–542, May. 2018.

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

[8] S. Cui, Y. Wang, S. Wang, R. Wang, W. Wang, and M. Tan, “Real-time perception and positioning for creature picking of an underwater vehicle,” *IEEE Transactions on Vehicular Techology*, DOI: 10.1109/TVT.2020.2973656, 2020.

[9] M. Stachowsky, T. Hummel, M. Moussa, and H. A. Abdullah, “A slip detection and correction strategy for precision robot grasping,” *IEEE/ASME Transactions on Mechatronics*, vol. 21, no. 5, pp. 2214–2226, Apr. 2016.

[10] T. P. Tomo, S. Somlor, A. Schmitz, S. Hashimoto, S. Sugano, and L. Jamone, “Development of a hall-effect based skin sensor,” in *2015 IEEE SENSORS*, Busan, South Korea, Nov. 2015, pp. 1–4.

[11] Y. Bekiroglu, D. Kragic, and V. Kyrki, “Learning grasp stability based on tactile data and hmms,” in *19th International Symposium in Robot and Human Interactive Communication*, Viareggio, Italy, Sep. 2010, pp. 132–137.

[12] Y. Bekiroglu, K. Huebner, and D. Kragic, “Integrating grasp planning with online stability assessment using tactile sensing,” in *2011 IEEE International Conference on Robotics and Automation*, Shanghai, China, May. 2011, pp. 4750–4755.

[13] Y. Bekiroglu, D. Song, L. Wang, and D. Kragic, “A probabilistic framework for task-oriented grasp stability assessment,” in *2013 IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May. 2013, pp. 3040–3047.

[14] Y. Chebotar, K. Hausman, Z. Su, G. S. Sukhatme, and S. Schaal, “Self-supervised regrasping using spatio-temporal tactile features and reinforcement learning,” in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Deajeon, South Korea, Oct. 2016, pp. 1960–1966.

[15] B. S. Zapata-Impata, P. Gil, and F. Torres, “Non-matrix tactile sensors: How can be exploited their local connectivity for predicting grasp stability?,” *arXiv preprint arXiv:1809.05551*, 2018.

[16] A. Garcia-Garcia, B. S. Zapata-Impata, S. Orts-Escolano, P. Gil, and J. Garcia-Rodriguez, “Tactilegcn: A graph convolutional network for predicting grasp stability with tactile sensors,” *arXiv preprint arXiv:1901.06181*, 2019.

[17] 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?,” *arXiv preprint arXiv:1710.05512*, 2017.

[18] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, “Deep learning for tactile understanding from visual and haptic data,” in *2016 IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, May. 2016, pp. 536–543.

[19] H. P. Liu, Y. L. Yu, F. C. Sun, and J. Gu, “Visual-Tactile Fusion for Object Recognition,” *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 996-1008, Apr. 2017.

[20] S. X. Wang et al., “3D Shape Perception from Monocular Vision, Touch, and Shape Priors,” *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, Oct. 2018, pp. 1606-1613.

[21] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, “More than a feeling: Learning to grasp and regrasp using vision and touch,” *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300-3307, Jul. 2018.

[22] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah *et al.*, “Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks,” *arXiv preprint arXiv:1810.10191*, 2018.

[23] S. K. Kumar, “On weight initialization in deep neural networks,” *arXiv preprint arXiv:1704.08863*, 2017.

[24] P. De Boer, D. P. Kroese, Owens, S. Mannor, and R. Y. Rubinstein, “A tutorial on the cross-entropy method,” *Annals of operations research*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[25] D. Kingma, and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.