Generalized Visual-Tactile Transformer Network for Slip Detection

Shaowei Cui [*,**] **Junhang Wei** [*,**] **Xiaocan Li** [*] **Rui Wang** [*]
Yu Wang [*] **Shuo Wang** [*,***]

[*] *The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190 China*
(corresponding author: Shuo Wang shuo.wang@ia.ac.cn*).*
[**] *School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China*
[***] *Center for Excellence in Brain Science and Intelligence Technology Chinese Academy of Sciences, Shanghai 200031, China.*

Abstract: Slip detection plays a vital role in robotic dexterous grasping and manipulation, and it has long been a challenging problem in the robotic community. Different from traditional tactile perception-based methods, we propose a Generalized Visual-Tactile Transformer (GVT-Transformer) network to detect slip based on visual and tactile spatiotemporal sequences. The main novelty of GVT-Transformer is its ability to address unaligned vision and tactile data in various formats captured by various tactile sensors. Furthermore, we train and test our proposed network on a public and our visual-tactile grasping datasets. The experimental results show that our method is more suitable for sliding detection tasks than previous visual-tactile learning methods and more versatile.

Keywords: Information and sensor fusion, Perception and sensing, Intelligent robotics, Deep neural networks, Visual-tactile fusion perception.

1. INTRODUCTION

With the rapid development of representation learning methods (see Sünderhauf et al. (2018)) and robot learning methods (see Kroemer et al. (2019)), the perception and decision-making ability of the robots have been improved rapidly, and it has been possible to solve Rubik's cube with one hand (see OpenAI et al. (2019)), play Jenga (see Fazeli et al. (2019)) and perform other complex manipulation tasks (see Sanchez et al. (2018)). However, there are still some fundamental but challenging issues in robotic grasping and manipulating tasks. For grasping, it is necessary to manage the force distribution between the object and gripper before lifting and during grasping to ensure grasping stability (see Stachowsky et al. (2016)). Manipulation additionally deals with the contact dynamics between objects and the gripper during executing desired motions or undergoing changes form the external environment (see Wu et al. (2019)). For these issues, the primary issue is how to adjust actions to respond to changing contacts. Detecting slip and incipient slip can assist robots to automatically adjust the grasping force may provide a solution to this problem (see Su et al. (2015)). It is not surprising that many studies have developed various tactile sensors and corresponding learning methods to address this issue in the past decades (see Yousef et al. (2011), Van Wyk and Falco (2018)). Some recent review papers about slip detection and robotic tactile perception can be found in Francomano et al. (2013) and Luo et al. (2017).

Humans can naturally estimate the grasping force required to lift an unknown object and grasp it and can adjust the force accordingly. In this process, humans mainly rely on the sensory-sensitive tactile afferents (FA-I, SA-I, FA-II, and SA-II) with a reasonable distribution and excellent neural processing system (see Yousef et al. (2011)). Furthermore, vision also plays a critical role in this determination process, especially when the performance of current tactile sensors is far less than that of humans.

Recently, robotic visual-tactile perception has successfully been used for a variety of tasks, such as surface classification (see Gao et al. (2016)), object recognition (see Liu et al. (2017)), contact-rich tasks (see Lee et al. (2019)), etc. These studies strongly suggest that visual-tactile perception has better performance than visual-only or tactile-only perception. Unfortunately, most of the existing studies fuse the visual and tactile data by Early Fusion (EF) methods that have been shown to the lack of ability to capture modal-specific and cross-modal features in other multimodal sequence learning tasks (see Li et al. (2018), Zadeh et al. (2018)). Additionally, the existing methods are designed for tasks with specific tactile data format (e.g., image, matrices, vector, etc.) and aligned sequences, which leads to significant limitations on these platform generalization performance.

[⋆] This work was supported in part by the National Key R&D Program of China Under Grant No. 2018AAA0103003 and the National Natural Science Foundation of China under Grant 61773378.

To tackle these issues, we propose a Generalized Visual-Tactile Transformer network (GVT-Transformer) to learn features from aligned or unaligned visual-tactile sequences for slip detection. Our primary contributions are:

- A representation learning method from which appropriate visual-tactile features can be learned for a slip detection task.
- We demonstrate that the proposed method is more suitable for sliding detection tasks than the traditional early fusion method.
- We show that the generalization of our proposed model for tactile sensors with different data formats and whether the sequences from the two modalities are aligned.

2. RELATED WORK

2.1 Slip Detection

Slip detection has always been one of the hot research fields in robot manipulation community. Yousef et al. (2011) presents a robotic grasp controller that allows a parallel jaw gripper to gently pick up and set down unknown objects once a grasp location has been selected. The controller selects an appropriate initial grasping force, detects whether an object is slipping from the grasp, increases the grasp force as needed, and determines when to release an object to set it down. Later, Stachowsky et al. (2016) proposes a slip detection and correction strategy for precision robot grasping by a common matrix tactile sensor. More recently, Dong et al. (2017) detect slip using a GelSight sensor by directly measuring the relative displacement between the marker and texture for textured objects. Zhang et al. (2018) and Zapata-Impata et al. (2019) both adopt CNN with ConvLSTM networks to learn spatiotemporal tactile features for slip detection, and the tactile signal is captured by BioTac and FingerVision sensors, respectively. Unfortunately, tactile-only perception may not achieve the desired detection performance due to the limitation of sensing capability.

Adding visual perception is an intuitive solution. Li et al. (2018) proposes a visual-tactile learning method based on a deep neural network to detect slip, which shows the importance of visual-tactile fusion perception in slip detection tasks. A compact and multimodal representation of the sensory inputs was learned by self-supervision (see Calandra et al. (2018)), which can be used to improve the sample efficiency of policy learning. However, these methods concatenate features from visual and tactile modalities directly, which called early-fusion and have been shown the lack of the ability to capture modal-specific and cross-modal features (see Zadeh et al. (2018)). Additionally, most of the above methods are explicitly designed for tactile sensors with different data formats and sample rates, which leads to significant limitations in platform generalization performance.

2.2 Transformer Network

Vaswani et al. (2017) introduces the transformer network for neural machine translation tasks firstly in 2017. It consists of an encoder and a decoder constructed with the self-attention mechanism (see Parikh et al. (2016)), and shows powerful temporal information capture capabilities from the source domain to the target domain. In addition, transformer networks have also been successfully applied to many other tasks (see Strubell et al. (2019)). More recently, Tsai et al. (2019) proposes a Multimodal Transformer to model multimodal human language time-serials without explicitly aligning the data.

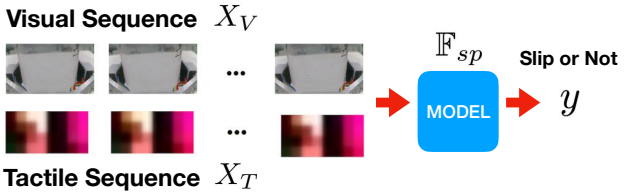

Fig. 1. A framework of slip detection using visual-tactile perception.

We absorb strong inspirations from the multimodal transformer to extend to a visual-tactile setting. Visual-tactile learning is different from conventional multimodal learning in that it is more focused on the capture of cross-modal and cross-temporal interaction features. It pays more attention to the ability to extract the fusion features of two modalities simultaneously, which is not implemented in Tsai et al. (2019).

3. TASK FORMULATION

We define the slip detection task as a binary classification problem, as shown in Fig. 1. Given the visual X_V and tactile X_T sequences, the slip detector model $\mathbb{F}_{sp}$ output the detection result y (slip or not). This task is formulated as

$$\begin{aligned} y &= \mathbb{F}_{sp}(X_V, X_T) \\ X_V &= \{x_v^0, x_v^1, ..., x_v^{T_V}\} \\ X_T &= \{x_t^0, x_t^1, ..., x_v^{T_T}\} \end{aligned} \tag{1}$$

where $y = 0\, or\, 1$, which $y = 0$ denotes a slip occurs at the current moment. x_v^0 and x_t^0 represent the first element in the visual and tactile sequences, respectively. T_V and T_T indicate the length of two sequences. Note that T_V and T_T are usually not equal due to the different sample rates of visual and tactile sensors, which means the model $\mathbb{F}_{sp}$ may need to address unaligned sequences from the two modalities.

4. PROPOSED METHOD

The overall architecture of GVT-Transformer is shown in Fig. 2. Firstly, the feature extraction modules and temporal convolution layers are used to extract visual and tactile features. Next, the modal-specific and cross-modal features are extracted by modal-specific and cross-modal transformers, respectively. Then, the modal-specific and cross-modal features of each modal are concatenated and fed into a fusion transformer layer. Finally, the final tactile and visual features are concatenated and sent into a classification layer to obtain an output. The detailed description of each component in GVT-Transformer is shown in this section.

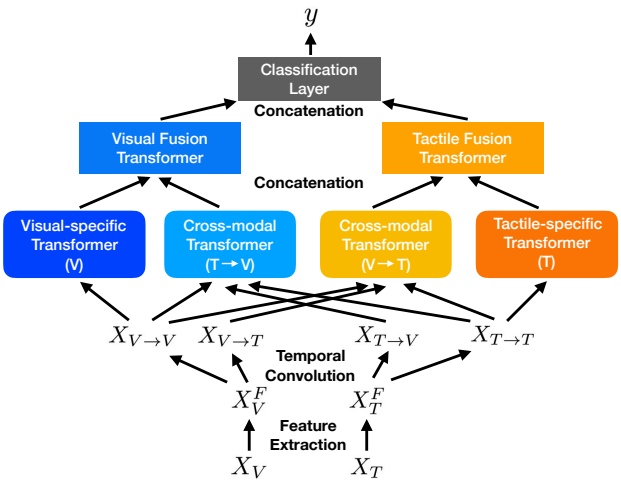

Fig. 2. The diagram of GVT-Transformer model.

4.1 Feature Extraction

Given the visual sequence ($X_V \in \mathbb{R}^{T_V \times D_V}$) and tactile sequence ($X_T \in \mathbb{R}^{T_T \times D_T}$), we first extract visual features $X_V^F \in \mathbb{R}^{T_V \times d_V}$ and tactile features $X_T^F \in \mathbb{R}^{T_T \times d_T}$, respectively. The visual features are extracted by pretrain Inception-v3 convolution neural network (CNN) (see Li et al. (2018)). The tactile features are extracted by different methods for different data formats, including LSTM, WaveNet (see Lee et al. (2019)), CNN, etc.

$$\begin{aligned} X_V^F &= \mathbb{F}_f^V \in \mathbb{R}^{T_V \times d_V} \\ X_T^F &= \mathbb{F}_f^T \in \mathbb{R}^{T_T \times d_T} \end{aligned} \tag{2}$$

where d_V and d_T indicate the dimensions of visual and tactile features, respectively.

4.2 Temporal Convolutions

To ensure that each element of the input sequences has sufficient awareness of its neighborhood elements, we encode the input sequences through some 1D temporal convolution layers:

$$\begin{aligned} X_{V\to V} &= \text{Conv1D}(X_V^F) \in \mathbb{R}^{T_V \times d_V^C} \\ X_{V\to T} &= \text{Conv1D}(X_V^F) \in \mathbb{R}^{T_V \times d_T^C} \\ X_{T\to T} &= \text{Conv1D}(X_T^F) \in \mathbb{R}^{T_T \times d_T^C} \\ X_{T\to V} &= \text{Conv1D}(X_T^F) \in \mathbb{R}^{T_T \times d_V^C} \end{aligned} \tag{3}$$

where d_V^C and d_T^C indicate the dimension of visual and tactile features after temporal convolutions, respectively. Note that $X_{V\to V}$ and $X_{T\to T}$ are used to extract modal-specific features, while $X_{V\to T}$ and $X_{T\to V}$ are used to extract cross-modal features.

4.3 Modal-specific Transformers

Different form Tsai et al. (2019), we build modal-specific transformers to extract modal-specific features further. In this subsection, we introduce the Tactile-Specific Transformer (TST) as an example to explain the modal-specific transformers. Given the tactile features $X_{T\to T}$, the tactile-specific transformer first adds position information on it by Positional Embedding (PE) to obtain $X_{T\to T}^{PE}$, which is

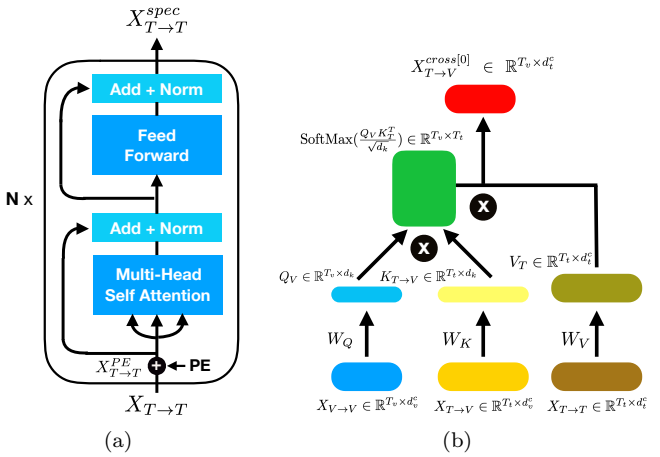

Fig. 3. (a) The tactile-specific transformer. (b) The Cross-modal attention module ($T \to V$).

fed into $N\times$ self-attention transformer layers to compute tactile-specific features $X_{T\to T}^{spec}$, as shown in Fig. 3(a). For specific details of the PE and self-attention transformer layers, please refer to Vaswani et al. (2017). Finally, we get the tactile-specific features

$$X_{T\to T}^{spec} = \text{TST}(X_{T\to T}) \in \mathbb{R}^{T_T \times d_T^C} \tag{4}$$

Note that the modal-specific transformers do not change the shape of the input feature maps.

4.4 Cross-modal Transformers

Similar to Tsai et al. (2019), we also build cross-modal transformers to extract cross-modal features in visual-tactile learning. Also, the biggest difference between the module constructed in this paper and Tsai et al. (2019) is that the number of input feature channels of different modalities is not necessarily consistent, which ensures that our proposed model can address different tactile data. The detailed architecture of the Cross-Modal Attention module ($\text{CMA}_{T\to V}$) is shown in Fig. 3(b), which is the core module of the proposed cross-modal transformer ($T \to V$).

Given the features $X_{V\to V} \in \mathbb{R}^{T_V \times d_V^C}$, $X_{T\to V} \in \mathbb{R}^{T_T \times d_V^C}$, and $X_{T\to T} \in \mathbb{R}^{T_t \times d_T^C}$, $\text{CMA}_{T\to V}$ first obtains three features by

$$\begin{aligned} Q_V &= X_{V\to V} W_Q,\ W_Q \in \mathbb{R}^{d_V^C \times d_K} \\ K_{T\to V} &= X_{T\to V} W_K,\ W_K \in \mathbb{R}^{d_V^C \times d_K} \\ V_T &= X_{T\to T} W_V,\ W_V \in \mathbb{R}^{d_T^C \times d_T^C} \end{aligned} \tag{5}$$

where $Q_V \in \mathbb{R}^{T_V \times d_K}$, $K_{T\to V} \in \mathbb{R}^{T_T \times d_K}$, and $V_T \in \mathbb{R}^{T_T \times d_T^C}$. $d_K$is a scaled dimension of Q_V and $K_{T\to V}$, which is set as $\sqrt{d_V^C}$ in this paper. Next, the correlation matrix of all positions of the two modal sequences can be calculated by

$$\text{Corr}_{T\to V} = \text{SoftMax}(\frac{Q_V K_{T\to V}^T}{\sqrt{d_k}}) \tag{6}$$

where $\text{Corr} \in \mathbb{R}^{T_V \times T_T}$. Finally, the output of $\text{CMA}_{T\to V}$ is obtained by

$$X_{T\to V}^{cross[0]} = \text{Corr}_{T\to V} \times V_T \tag{7}$$

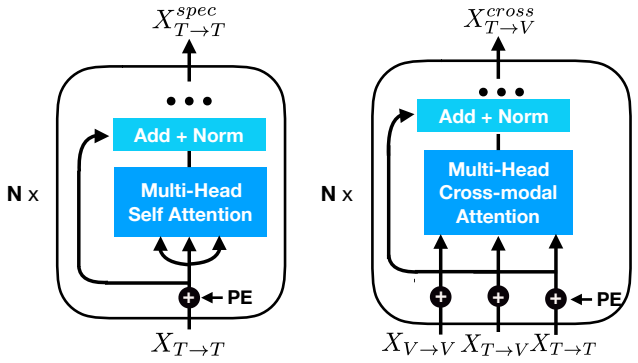

Fig. 4. The different inputs and attention module of cross-modal transformers and model-specific transformers.

where $X^{cross[0]}_{T\to V} \in \mathbb{R}^{T_V \times d^C_T}$. The superscripts [0] represents the cross-modal features obtained by inputs passing through the cross-modal attention module for the first time. In the next layers, $X_{V\to V}$ will be replaced by the output of the previous layer as the new input, while $X_{T\to V}$ and $X_{T\to T}$ will not.

$$X^{cross[i]}_{T\to V} = \mathrm{CMA}_{T\to V}(X^{cross[i-1]}_{T\to V}, X_{T\to V}, X_{T\to T}) \quad (8)$$

where $i > 0$ is the index of transformer layers.

In this way, the cross-modal transformer ($T \to V$) uses the $\mathrm{CMA}_{T\to V}$ instead of self-attention module in model-specific transformers to complete the cross-modal features ($T \to V$) extraction. The different key parts of cross-modal and modal-specific transformer are shown in Fig. 4.

4.5 Fusion Transformers and the Classification Module

After the modal-specific and cross-modal features extraction, we concatenate these features and sent them into visual and tactile fusion transformers to extract the final feature of each modal. The architecture of these fusion transformers are the same as the model-specific transformer, as shown in Fig. 3(a).

Finally, the extracted final visual ($X^R_V \in \mathbb{R}^{T_V \times (d^C_V + d^C_T)}$) and tactile ($X^R_T \in \mathbb{R}^{T_T \times (d^C_V + d^C_T)}$) features are concatenated and fed into the fully connected (FC) layers for classification, and the final classification result y is obtained by

$$y = \mathrm{FC}((X^R_V \oplus X^R_T)) \quad (9)$$

where $y = 0\, or\, 1$.

5. EXPERIMENTS AND ANALYSIS

In this section, we perform comparative experiments on two visual-tactile slip detection datasets to verify the performance of the proposed GVT-Transformer model. Our goal is to answer the following two questions:

(1) Is our proposed model more suitable for slip detection tasks than the Early Fusion (EF) method?
(2) Can the proposed model accommodate unaligned visual and tactile sequence data and different tactile sensors?

Fig. 5. The visual and tactile sequences of **D1**.

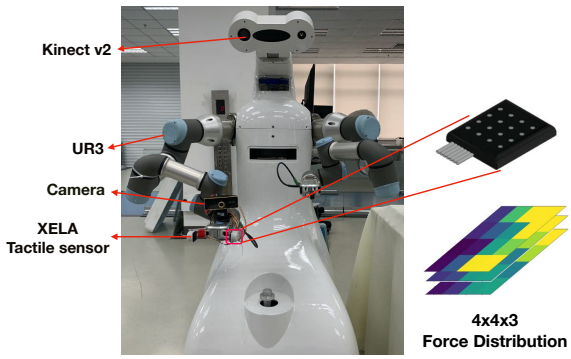

Fig. 6. The grasping setup of **D1**.

5.1 Visual-Tactile Grasping Datasets Introduction

D0: Li et al. (2018) build a visual-tactile grasping dataset for slip detection tasks, in which the training data is acquired by a GelSight tactile sensor (see Yuan et al. (2017)), and a camera mounted on the gripper, respectively. Note that the train-test partition of **D0** is not completely consistent with the original paper.

D1: We collect visual and tactile sequence data by two XELA tactile sensors (see Tomo et al. (2015)) and a camera mounted on a gripper, as shown in Fig. 6. An example of collected visual-tactile sequences is shown in Fig. 5. In total, we perform 5,000 grasps on 17 dairy objects, and the number of grasping routines is nearly 10 times as **D0** did. For more detailed information of **D1**, please refer to our detailed dataset [1].

Note that the visual-tactile sequences of **D0** are aligned, while **D1** is not.

*5.2 Experiments on **D0***

Implementation details: We use *Inception-V3* as the backbone CNN to extract features, and the sequence length is set to 8 according to Li et al. (2018). The detailed parameters of GVT-Transformer on dataset **D0** is shown in Table 1. We use Xavier initialization, cross-entropy loss function, Adam optimizer with a learning rate of 5e-5, and a batch size of 64 to train the models on an Nvidia GeForce Titan X GPU platform with PyTorch 1.3.0 package.

Results: The precision, recall, and F1 score of the two models on **D0** are shown in Table 2. The detection performance of the proposed method is not inferior and even better than EF in an aligned situation, which answers our first question.

[1] `https://drive.google.com/drive/folders/1IcYp4oIjFWUlw-X8Ei27JSBiaeiV32Ts?usp=sharing`

Table 1. Details network parameters of GVT-Transformer on **D0**

Layers	Operations	FeatureMap
T/V CNN	Inception v3(avg-pool)	8 × 2048
T/V-Conv1D	outdim(64),kernel(2),padding(0)	7 × 64
T/V-S TF	layers(3),head(4)	7 × 64
T/V-C TF	layers(3),head(4)	7 × 64
T/V-B TF	layers(5),head(4)	7 × 128
Classification	FC(256,64),FC(64,2)	2

Table 2. Comparison of classification performance on **D0**

Methods	Precision (%)	Recall (%)	F1 score (%)
Early Fusion	84.75	84.75	84.74
GVT-Transformer	**85.83**	**85.37**	**85.28**

*5.3 Experiments on **D1***

Implementation details: We set the initial network parameters of GVT-Transformer shown in Table 3, and the detailed parameter optimization process can be found in the supplementary material [2]. Note that the training strategy here is the same as experiments on **D0** except for batch size (512) and learning rate (1e-6).

Table 3. Initial parameters of proposed model

Layers	Operations	Feature Map
V CNN	Inception v3(avg-pool)	5 × 2048
T Signal	No operations	11 × 92
T/V Conv1D	out(64/64),kernel(2),pad(0)	10 × 32/4 × 64
T/V-S TF	layers(5),head(8)	10 × 32/4 × 64
T/V-C TF	layers(5),head(8)	10 × 64/4 × 32
T/V-F TF	layers(5),head(8)	10 × 96/4 × 96
Classification	FC(192,64),FC(64,2)	2

Ablation study: To further study the influence of the individual components in GVT-Transformer, we perform a comprehensive ablation analysis on dataset **D1**. The results are shown in Table 4.

Table 4. An ablation study of GVT-Transformer on **D0**.

Parameters	Precision (%)	Recall (%)	F1 score (%)
Visual-only	72.67	58.66	61.39
Tactile-only	96.83	89.61	92.30
No modal-spec	85.58	89.69	87.48
No cross-modal	83.39	69.99	74.74
GVT-Transformer	**97.43**	90.12	**93.37**
Adapted EF	90.83	**90.89**	92.50

Firstly, we consider the performance for only using unimodal transformers, i.e., tactile or visual only. The experiment results show that the tactile-only transformer outperforms the visual-only transformers, which indicates that tactile plays a more critical role in the slip detection task. This finding aligns with the observations in prior work (see Li et al. (2018)). Furthermore, we consider the modal-specific transformers and cross-modal transformers. The experiment results show both of them are useful for the task, and cross-modal features are even more effective,

[2] `https://drive.google.com/file/d/1XoWQMfucB2bMf6jpzFzIqOYrZ8gQ40CM/view?usp=sharing`

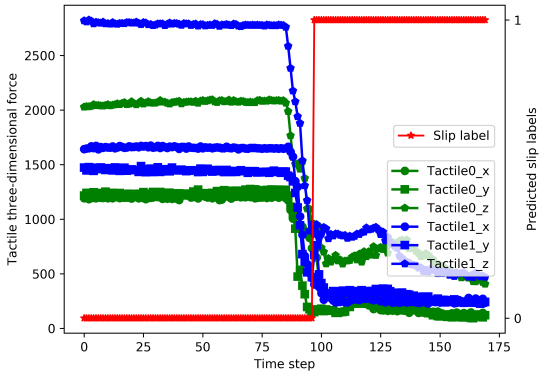

Fig. 7. Three-dimensional force readings for two tactile sensors and a slip detection label variation plot. Predicted slip label: 0 means no slip and 1 denotes slip.

which also indicates that the visual-tactile improvements based on Tsai et al. (2019) are beneficial for this task.

Furthermore, we also apply the adapted Early Fusion method on **D1** by directly concatenating tactile data to align with visual sequence, and the experimental results show that our method outperforms adapted EF by a large margin (about 7% precision). In other words, GVT-Transformer is better suited for addressing unaligned visual-tactile data, which answers our second question.

5.4 Experiments on robots

A slip detection experiment based on GVT-Transformer is performed on our robot to verify its effectiveness in practice. A Nestlé bottle with a half bottle of water is selected as a grasping object. We first set the initial grasping force of 15N, which is sufficient to hold and hold the bottle stably. During the lifting process, we suddenly set the grasping width to 71 mm so that a slip will occur without causing the water bottle to fall. In this grasping routine, the three-dimensional force readings of two XELA tactile sensors and the slip detection labels changes are shown in Fig. 7. Obviously, the label changes immediately following the command and the grasping video can be found in `https://youtu.be/oGkhwo9yGMQ`.

6. CONCLUSION

A transformer network-based slip detection method named GVT-Transformer is proposed to solve the unaligned visual-tactile sequence learning problem in this paper. Specifically, modal-specific and cross-modal transformers of GVT-Transformer are presented to capture modal-specific and cross-modal features, respectively. Furthermore, The experimental results show that our proposed method not only outperforms traditional early fusion methods in detection performance but also more suitable for unaligned situation. A sliding detection experiment carried out on a real robot also confirm the proposed method.

REFERENCES

Calandra, R., Owens, A., Jayaraman, D., Lin, J., Yuan, W., Malik, J., Adelson, E.H., and Levine, S. (2018). More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4), 3300–3307. doi:10.1109/LRA.2018.2852779.

Dong, S., Yuan, W., and Adelson, E.H. (2017). Improved gelsight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 137–144. IEEE.

Fazeli, N., Oller, M., Wu, J., Wu, Z., Tenenbaum, J.B., and Rodriguez, A. (2019). See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion. *Science Robotics*, 4(26), eaav3123. doi:10.1126/scirobotics.aav3123.

Francomano, M.T., Accoto, D., and Guglielmelli, E. (2013). Artificial Sense of SlipA Review. *IEEE Sensors Journal*, 13(7), 2489–2498. doi:10.1109/JSEN.2013.2252890. URL `http://ieeexplore.ieee.org/document/6479676/`.

Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., and Darrell, T. (2016). Deep learning for tactile understanding from visual and haptic data. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June, 536–543. doi:10.1109/ICRA.2016.7487176.

Kroemer, O., Niekum, S., and Konidaris, G. (2019). A Review of Robot Learning for Manipulation: Challenges, Representations, and Algorithms. URL `http://arxiv.org/abs/1907.03146`.

Lee, M.A., Zhu, Y., Srinivasan, K., Shah, P., Savarese, S., Fei-Fei, L., Garg, A., and Bohg, J. (2019). Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. 8943–8950. doi:10.1109/icra.2019.8793485.

Li, J., Dong, S., and Adelson, E. (2018). Slip Detection with Combined Tactile and Visual Information. *Proceedings - IEEE International Conference on Robotics and Automation*, 7772–7777. doi:10.1109/ICRA.2018.8460495.

Liu, H., Yu, Y., Sun, F., and Gu, J. (2017). Visual-Tactile Fusion for Object Recognition. *IEEE Transactions on Automation Science and Engineering*. doi:10.1109/TASE.2016.2549552.

Luo, S., Bimbo, J., Dahiya, R., and Liu, H. (2017). Robotic tactile perception of object properties: A review. *Mechatronics*, 48, 54–67.

OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., and Zhang, L. (2019). Solving Rubik's Cube with a Robot Hand. 1–51. URL `http://arxiv.org/abs/1910.07113`.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. (2016). A Decomposable Attention Model for Natural Language Inference. doi:10.18653/v1/d16-1244.

Sanchez, J., Corrales, J.A., Bouzgarrou, B.C., and Mezouar, Y. (2018). Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. *International Journal of Robotics Research*, 37(7), 688–716. doi:10.1177/0278364918779698.

Stachowsky, M., Hummel, T., Moussa, M., and Abdullah, H.A. (2016). A Slip Detection and Correction Strategy for Precision Robot Grasping. *IEEE/ASME Transactions on Mechatronics*, 21(5), 2214–2226. doi:10.1109/TMECH.2016.2551557.

Strubell, E., Verga, P., Andor, D., Weiss, D., and McCallum, A. (2019). Linguistically-Informed Self-Attention for Semantic Role Labeling. doi:10.18653/v1/d18-1548.

Su, Z., Hausman, K., Chebotar, Y., Molchanov, A., Loeb, G.E., Sukhatme, G.S., and Schaal, S. (2015). Force estimation and slip detection/classification for grip control using a biomimetic tactile sensor. *IEEE-RAS International Conference on Humanoid Robots*, 2015-Decem, 297–303. doi:10.1109/HUMANOIDS.2015.7363558.

Sünderhauf, N., Brock, O., Scheirer, W., Hadsell, R., Fox, D., Leitner, J., Upcroft, B., Abbeel, P., Burgard, W., Milford, M., et al. (2018). The limits and potentials of deep learning for robotics. *The International Journal of Robotics Research*, 37(4-5), 405–420.

Tomo, T.P., Somlor, S., Schmitz, A., Hashimoto, S., Sugano, S., and Jamone, L. (2015). Development of a hall-effect based skin sensor. In *2015 IEEE SENSORS - Proceedings*. doi:10.1109/ICSENS.2015.7370435.

Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., and Salakhutdinov, R. (2019). Multimodal Transformer for Unaligned Multimodal Language Sequences. 6558–6569. doi:10.18653/v1/p19-1656.

Van Wyk, K. and Falco, J. (2018). Slip Detection: Analysis and Calibration of Univariate Tactile Signals. URL `http://arxiv.org/abs/1806.10451`.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), 5999–6009.

Wu, Y., Yan, W., Kurutach, T., Pinto, L., and Abbeel, P. (2019). Learning to Manipulate Deformable Objects without Demonstrations. URL `http://arxiv.org/abs/1910.13439`.

Yousef, H., Boukallel, M., and Althoefer, K. (2011). Tactile sensing for dexterous in-hand manipulation in robotics - A review. *Sensors and Actuators, A: Physical*, 167(2), 171–187. doi:10.1016/j.sna.2011.02.038. URL `http://dx.doi.org/10.1016/j.sna.2011.02.038`.

Yuan, W., Dong, S., and Adelson, E.H. (2017). GelSight: High-resolution robot tactile sensors for estimating geometry and force. doi:10.3390/s17122762.

Zadeh, A., Vij, P., Liang, P.P., Cambria, E., Poria, S., and Morency, L.P. (2018). Multi-attention recurrent network for human communication comprehension. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 5642–5649.

Zapata-Impata, B.S., Gil, P., and Torres, F. (2019). Learning Spatio temporal tactile features with a convLSTM for the direction of slip detection. *Sensors (Switzerland)*, 19(3), 1–16. doi:10.3390/s19030523.

Zhang, Y., Kan, Z., Tse, Y.A., Yang, Y., and Wang, M.Y. (2018). FingerVision Tactile Sensor Design and Slip Detection Using Convolutional LSTM Network. *ArXIV*. URL `http://arxiv.org/abs/1810.02653`.