信息与控制　2024 年　第 53 卷　第 2 期：191 ~ 198

DOI：10.13976/j.cnki.xk.2023.2598　　文章编号：1002 - 0411(2024) - 03 - 0191 - 08

基于注意力机制和视触融合的机器人抓取滑动检测

黄兆基[1]，高军礼[1]，唐兆年[1]，宋海涛[2]，郭　靖[1]

1. 广东工业大学自动化学院，广州　510006；　2. 华南理工大学工商管理学院，广州　510640

基金项目：国家自然科学基金项目(61803103)；国家留学基金项目(201908440537)

通信作者：高军礼，jonygao621@gdut.edu.cn　收稿/录用/修回：2022 - 12 - 29/2023 - 02 - 23/2023 - 04 - 08

摘要

机器人抓取滑动检测对于抓取任务而言具有重要的意义。在抓取过程中，视觉和触觉作为判断抓取状态的关键模态信息，实现其高效融合仍具有挑战性。基于视触觉模态信息融合理念，提出一种新型视触觉融合模型，用于高效解决机器人抓取滑动检测问题。首先，该模型通过卷积神经网络、多尺度时序卷积网络提取视触觉数据的空间和时序特征信息；然后，使用注意力机制为视触觉特征分配权重，并通过多尺度时序卷积网络进行多模态信息融合；最后，通过全连接层输出机器人抓取状态的检测结果。使用 7 自由度 XArm 机械臂、D455 RGB 摄像头和 XELA 触觉传感器进行数据采集。实验结果表明，基于该模型的机器人抓取滑动检测的准确率高达 98.98%，该模型在可靠、顺利执行机器人抓取任务方面具有较好的研究与应用价值。

关键词

机器人抓取

视触融合

多尺度时序卷积网络

注意力机制

中图法分类号：TP242.6

文献标志码：A

Slip Detection for Robot Grasping Based on Attention Mechanism and Visuo-Tactile Fusion

HUANG Zhaoji[1], GAO Junli[1], TANG Zhaonian[1], SONG Haitao[2], GUO Jing[1]

1. *School of Automation*, *Guangdong University of Technology*, *Guangzhou* 510006, *China*;

2. *School of Business Administration*, *South China University of Technology*, *Guangzhou* 510640, *China*

Abstract

Slip detection for robot grasping is of great significance for grip tasks. In the grasping process, vision and touch are the key modal information for judging the grasping state. To achieve efficient visuotactile fusion remains challenging. A novel visual and tactile fusion model is proposed to address the issues of efficient slip detection for robot grasping based on the idea of visual and tactile fusion. First, the model extracts the spatial and temporal feature information of visual and haptic data using a convolutional neural network and a multiscale temporal convolutional network. Second, an attention mechanism is used to assign weights to visuotactile features, and multimodal information fusion is performed through a multiscale temporal convolutional network. Finally, the detection results of the grasping state are acquired through one fully connected layer. Data acquisition is implemented using an XArm 7DOF robotic arm, D455 RGB camera, and XELA tactile sensor. The experimental results show that the accuracy rate of slip detection for robot grasping using the proposed model is as high as 9898%. The proposed model has good research and application value in the reliable and smooth execution of robot grip tasks.

Keywords

robot grasp;

visuo-tactile fusion;

multi-scale temporal convolution network;

attention mechanism

0 引言

机器人在执行抓取任务时，对抓取状态进行准确检测是任务成功执行的关键。当检测到滑动发生时，机器人需要及时调整抓取力度和抓取策略，以实现稳定抓取[1]。要实现高效地滑动检测，充分利用各种模态的信息是关键。视觉模态和触觉模态是滑动检测领域中重要的两种模态信息。由于视觉模态信息和触觉模态信息在形式和结构上的不同，实现视触觉模态信息的高效特征提取和融合仍是一个巨大的挑战。

得益于触觉传感器的快速发展[2-3]，触觉信息在机器人抓取滑动检测领域的应用得到广泛关注[4]。JAMES 等[5]使用支持向量机检测机器人抓取时的滑动状态，当有滑动发生时及时调整抓取策略，使物体能够被重新稳定地抓取。GRIFFA 等[6]提出一种基于模型的滑动检测方法，实时预测机器人可能失败的抓握并及时增加抓握力以实现稳定抓取。SUI 等[7]使用基于视觉的触觉传感器采集数据，提出一种基于分布力和变形梯度的初始滑动检测方法。针对非矩阵形式的触觉传感器，GARCIA-GARCIA 等[8]使用图卷积神经网络(graph convolution neural networks，GCNNs)处理触觉数据，充分利用非矩阵触觉传感器感应点之间的位置信息。

人类在进行稳定抓取时，视觉和触觉反馈都起着重要作用[9]。视觉信息能够提供物体丰富的几何和位置信息。触觉信息能够提供接触时丰富的力反馈信息。CALANDRA 等[10]通过实验验证，同时使用视觉和触觉信息的机器人抓取滑动检测效果比只使用其中任何一种模态信息的效果都要好。LI 等[11]提出了一种基于长短期记忆网络(long short-term memory networks，LSTM)的视触觉融合滑动检测方法。通过预训练网络 Inception V3 提取视触觉数据的空间特征，并使用 LSTM 网络进行融合，最终得到 88.03% 的检测准确率。YAN 等[12]在文[11]所提出模型的基础上使用时序卷积网络(temporal convolutional network，TCN)替换 LSTM 网络来提取数据的时序特征和实现视触特征融合，这是 TCN 网络在机器人滑动检测领域的首次应用。通过实验证明，在相同条件下，使用 TCN 网络的模型的预测准确率相比于使用 LSTM 网络提高了 1.32%。CUI 等[13]针对相机和触觉传感器采样频率不相同的事实，通过 3D 卷积网络处理时间长度不一致的视觉数据和触觉数据。YAN 等[14]将机器人执行抬升动作前的稳定性预测和之后的滑移检测相结合，构建一个结合视觉和触觉的多阶段、多输出框架。

滑动抓取或稳定抓取是一个持续过程，其内蕴数据的时序特征有助于提升机器人抓取滑动检测的准确性。ZAPATA-IMPATA 等[15]对 LSTM 网络和 ConvLSTM网络在滑动检测中的性能进行测试和比较。结果显示，LSTM 的表现比 ConvLSTM 更优秀。在已知的滑动检测研究中，通常会使用循环神经网络(recurrent neural network，RNN)提取数据的时序特征。但 RNN 网络在处理长序列时容易丢失早期数据的必要特征信息。为此，LEA 等[16]在 2017 年提出时序卷积网络。通过 1 维卷积网络对输入做并行化处理，并使输出覆盖全部的输入，从而避免早期数据特征信息丢失的问题。随后，BAI 等[17]在多个时序相关的公开数据集上，综合比较 TCN、LSTM 和门控循环单元(gated recurrent unit，GRU)的性能。结果显示，TCN 的性能最好。MARTINEZ 等[18]在 TCN 网络的基础上，提出了多尺度时序卷积网络(multi-scale temporal convolution network，MS-TCN)并应用在唇语识别，在 LRW(Lip Reading in the Wild)[19]和 LRW1000[20]数据集上均有优秀的表现。MS-TCN 使 TCN 的性能得到进一步提升。鉴于 MS-TCN 网络在时序特征提取方面的良好性能，率先将 MS-TCN 网络用于提取视触觉数据的时序特征，并使用 MS-TCN 网络实现视触觉多模态信息的融合。

人类在抓取物体时所用到的视觉和触觉两种模态信息，在不同时刻的重要程度有所不同。例如，在黑暗环境下，触觉相比于视觉更重要；当抓取光滑物体时，视觉相比于触觉更重要。受此启发，在进行视触觉特征拼接时引入注意力机制[21]，对视触觉特征权重进行重新分配，使模型能够在抓取不同物体时均有良好表现。CG(context gate)[22]作为简单实用的门控单元，能够有效地给相关元素重新分配权重。为此，使用 CG 在视触觉特征融合之前重新为其分配权重。

在使用视觉和触觉信息进行滑动检测的研究中，主要是把触觉信息以图像处理的方式进行特征提取。文[11]通过基于视觉的触觉传感器采集高分辨率的触觉图片，使用预训练网络 Inception-V3 进行处理。文[13]中把通过 XELA 触觉传感器采集到的触觉数据转换为触觉图片，再使用卷积神经网络(convolutional neural network，CNN)提取其空间特征。该方法在把 XELA 触觉传感器采集到的原始数据映射到 0 ~ 255 时会容易丢失一些变化细微

的信息。例如，抓取过程中发生轻微的滑动，数据细微的变化映射到图片时就有可能丢失，造成检测错误。为此，提出一种基于触觉原始数据的空间特征提取模型，在不丢失细微变化信息的同时，实现对触觉数据的空间特征提取。

综上，本文提出一种结合注意力机制的视触觉融合的深度神经网络模型，用于机器人抓取物体时的滑动检测。对 50 个具有不同大小、形状、材料和重量的日常物体进行抓取和抬升动作的任务实验，并使用 XArm 7DoF 机械臂、D455 摄像头和 XELA 触觉传感器采集视觉数据和触觉数据。将其中 40 个物体的抓取数据作为训练集用于网络训练，余下 10 个物体的抓取数据作为测试集用于网络测试，在测试集上得到 98.98% 的检测准确率。此外，通过消融实验和对比实验进一步验证了模型的有效性。

1　视触觉融合模型

1.1　问题描述

本文所提出的视触觉融合模型结构如图 1 所示，两种输入模态分别是视觉输入序列 x_v^T 和触觉输入序列 x_t^T。x_v^T 和 x_t^T 分别经过特征提取模型 F_v 和触觉特征提取模型 F_t 得到对应的时空特征 T_v 和 T_t。将 T_v 和 T_t 进行拼接，并在拼接时使用注意力模块 F_a 为其进行权重分配。把拼接好的特征输入到视触觉融合模块 $F_{v,t}$，得到融合后的特征 $T_{v,t}$。最后，把 $T_{v,t}$ 输入到分类模型 F_c，得到一个二分类的结果 y。该过程的形式化描述如式(1)～式(4)所示：

$$T_v = F_v(x_v^T) \tag{1}$$

$$T_t = F_t(x_t^T) \tag{2}$$

$$T_{v,t} = F_{v,t}(F_a(T_v, T_t)) \tag{3}$$

$$y = F_c(T_{v,t}) \quad y \in (0, 1) \tag{4}$$

其中，T 表示输入序列的长度，输出 y 的值 0 和 1 分别表示滑动和稳定状态。

1.2　空间特征提取

网络层数越深，网络的表达能力越强。但对于一般的网络，较深的层数容易导致梯度消失的问题。ResNet 网络通过使用残差网络结构缓解了这个问题。通过残差网络结构，ResNet 网络可以堆叠很多层，并且网络的分类效果不会随着网络层数的增加而变差。残差网络的思想是：当一个网络已经达到了最优时，通过恒等映射来加深网络的深度并不会导致网络的误差变大，网络的性能也不会随着深度增加而降低。恒等映射就是输出等于输入。ResNet 网络通过串联多个残差块来实现深层网络，每个残差块由多个卷积层和恒等映射来组成。考虑到 ResNet 网络参数量较大，而本文采集的数据集相对较小，使用迁移学习的方法来训练网络。迁移学习的思想是将模型已训练好的权重参数迁移到当前模型中，使用当前数据集对模型进行微调，从而降低模型的训练难度。ResNet 网络经过 ImageNet 数据集训练后，具有很强的特征提取能力，广泛应用于迁移学习。常见的经过 ImageNet 数据集训练的预训练网络有 ResNet18、ResNet34、ResNet50、VGG-16、Inception-V3 等。不同的预训练网络针对同一个数据集，其表现也并不一样。经过对比实验，针对自主制作的数据集，ResNet34 预训练网络提取视觉数据空间特征的表现更好。

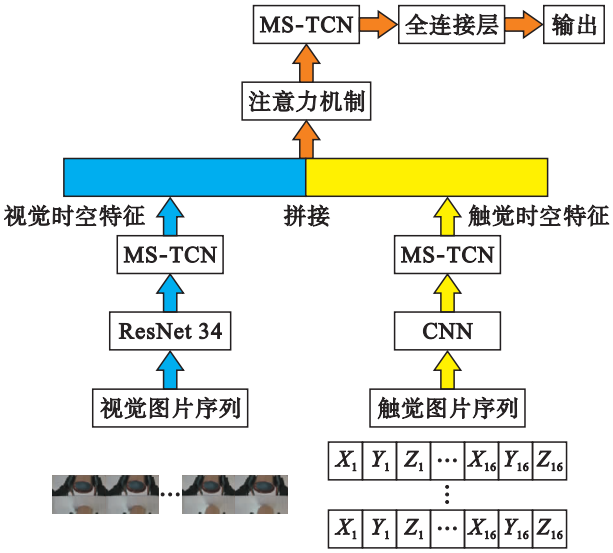

图 1　结合注意力机制的视触觉融合模型结构

Fig.1　The structure of the visuo-tactile fusion model combined with attention mechanism

对于触觉数据的采集，使用 XELA 触觉传感器。该传感器以 4 ×4 的形式分布着 16 个感应点，每个感应点可以检测 3D 力触觉数据，其中 Z 方向是垂直于传感器表面的法向力，X 和 Y 方向是平行于传感器表面两个方向的切向力。当物体沿 X 轴方向滑动时，Y 方向检测到的力触觉数据变化不会很大；当物体沿 Y 轴方向移动时，X 方向检测到的力触觉数据也不会有很大的变化；而当物体被稳定抓取时，X 和 Y 方向的力触觉数据都会相对稳定。考虑到这 3 个方向的力触觉数据具有不同的物理含义，在提取触觉数据的空间特征时，分别重组成与触觉传感器感应点一致的 4 ×4 的分布形式，再使用触觉特征提取模型分别提取 X、Y、Z 方向触觉数据的空间特征，最后再拼接在一起。触觉空间特征

提取模型结构和参数分别如图 2 和表 1 所示。触觉空间特征提取模型由 3 层卷积层组成，分别是 Conv 1、Conv 2 和 Conv 3。其中，Conv 1 由 3 个大小为 3 ×3 的卷积核组成，卷积核的步长为(1，1)，为了不改变输出特征图的尺寸，对输入进行填充；Conv 2 包含了 8 个大小为 3 ×3 的卷积核，卷积核的步长为(1，1)；Conv 3 拥有 16 个大小为 2 ×2 的卷积核，卷积核的步长为(1，1)。经过 3 层卷积层，每个方向的触觉数据可以分别得到 16 维的空间特征。对 3 个方向触觉数据的空间特征进行拼接，可以得到 48 维触觉数据空间特征。

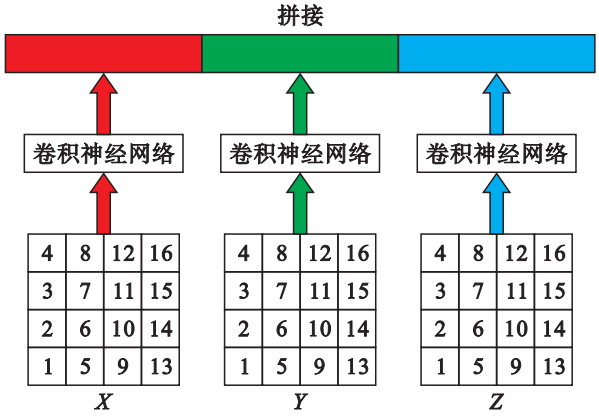

图 2 触觉空间特征提取模型

Fig.2 Haptic spatial feature extraction model

表 1 触觉空间特征提取模型参数

Tab.1 The parameters of haptic spatial feature extraction model

卷积层	参数	输出大小
Conv 1	3 ×3 ×3，padding(1，1)，stride(1，1)	4 ×4 ×3
Conv 2	3 ×3 ×8，stride(1，1)	2 ×2 ×8
Conv 3	2 ×2 ×16，stride(1，1)	1 ×1 ×16

注：stride 表示卷积核的步长，padding 表示填充大小。

1.3 时序特征提取

滑动是一个动态过程，因而对其内蕴含数据的时序特征的提取非常关键。在卷积神经网络中，1 维卷积网络的卷积核只会朝着一个方向移动，因此可以让卷积核沿着时间维度进行移动，从而提取数据之间的时序关系。1 维卷积网络的输入一般是一个 2 维矩阵，其中行表示某个时间节点数据的维度，列表示时间的长度。1 维卷积网络的卷积核要求在行数上与输入数据相同，卷积核的列数决定着每次卷积所覆盖的时间维度。空洞卷积表示卷积核所覆盖的元素之间存在一定的间距，用扩张率来表示，其作用是扩大卷积核的感受野。普通卷积是空洞卷积的一种特殊情况，普通卷积中元素之间的间距为 1。

MS-TCN 的每一层通过一维卷积网络将所有时间点的特征进行并行处理。层与层之间通过空洞卷积使用较少的隐藏层即可将顶层输出的感受野覆盖到全部输入。MS-TCN 的每一层还具有多个分支，且每个分支通过设定不同大小的卷积核使得每一层都能融合不同时间段的信息。为此，使用 MS-TCN 提取视触觉数据的时序特征。MS-TCN 的结构如图 3 所示，共包含两层，每层有两个分支，每个分支的卷积核大小都为 5。其中，Layer1 和 Layer2 分别表示第 1 层和第 2 层卷积层，branch1 _kernel 和 branch2_kernel 分别表示在某一层卷积层中分支 1 和分支 2 的卷积核大小。

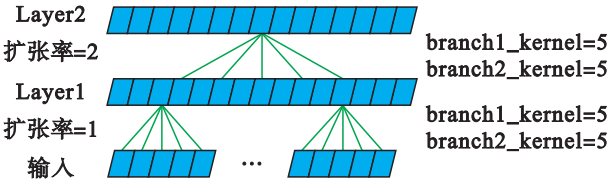

图 3 MS-TCN 结构图

Fig.3 The structure of MS-TCN

1.4 注意力和视触融合模块

不同环境下抓取不同物体时，视觉和触觉的重要程度并不相同。受此启发，在拼接视觉和触觉时空特征时，通过 CG 模块对视触觉时空特征重新分配权重。CG 表达式如式(5)所示：

$$\boldsymbol{Y}=\sigma(\boldsymbol{WX}+\boldsymbol{b})\boldsymbol{X} \tag{5}$$

其中，$\boldsymbol{X}\in\mathbb{R}^n$是输入特征向量，$\sigma$ 是 Sigmoid 激活函数，$\boldsymbol{W}\in\mathbb{R}^n$和 $\boldsymbol{b}\in\mathbb{R}^n$是可训练的参数。通过 σ 将 $\boldsymbol{WX}+\boldsymbol{b}$ 映射到[0，1]，表示输入 $\boldsymbol{X}$ 的权重。

通过 CG 模块得到重新分配权重并拼接在一起的视触觉特征，使用 MS-TCN 对其进行特征信息融合。该处的 MS-TCN 共有 3 层，每层有两个分支，每个分支的卷积核大小都为 3。视觉数据和触觉数据在经过特征提取后具有不同时间维度的特征。使用 MS-TCN 实现其特征信息融合，是利用 MS-TCN 的时序特征提取能力，在进行信息融合的同时考虑不同时间维度特征之间的关系。融合后的特征输入到分类模块，得到机器人抓取滑动的检测结果。分类模块由两层全连接层组成，最后一层全连接层输出二分类的结果。其中，0 和 1 分别表示滑动和稳定状态。

2 数据采集

使用XArm7 DoF机械臂以及配套的二指夹爪、

D455 摄像头和 XELA 触觉传感器所搭建的数据采集平台如图 4 所示。其中，XELA 触觉传感器贴装在夹爪的表面，D455 摄像头安装在夹爪上侧。

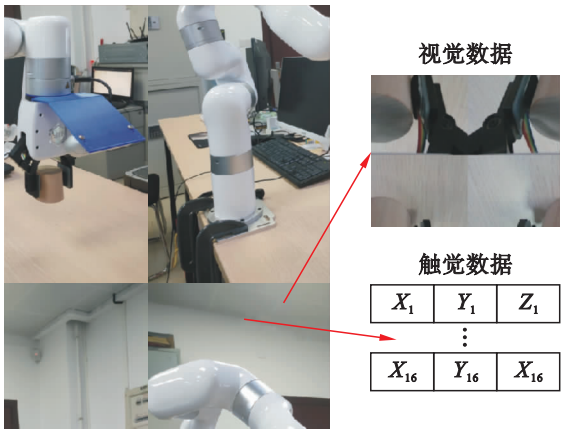

图 4　数据采集实验平台

Fig.4　The experimental platform for data acquisition

XELA 触觉传感器的感应点以 4 × 4 的形式分布，每个感应点可检测 X、Y、Z 三个维度的力触觉数据。D455 摄像头采集得到的 RGB 图片尺寸为 640 × 480 像素。选取 50 个具有不同大小、形状、材料和重量的日常物体进行数据采集，部分物体如图5所示。所选物体的宽度均小于夹爪的最大行程 77 mm。考虑到 XELA 触觉传感器的测量范围，物体的重量都小于 1 kg。数据集包含由 D455 摄像头和 XELA 触觉传感器分别采集得到的视觉和触觉数据。这两类数据在时间上是对应的，即同一时刻分别得到一张 RGB 图片和一份触觉数据。

在进行数据采集时，每个物体预先放在既定位置，机械臂会根据预先设定的抓取位置和抓取力度对物体执行抓取和抬升动作。数据采集在二指夹爪闭合完成的同时启动。二指夹爪闭合完成后，机械臂会向上抬升 30 mm，当机械臂抬升至顶端时关闭数据采集。随后，机械臂向下移动返回到抓取起始位置，松开夹爪，并执行下次抓取。机械臂对每个物体重复执行 15 ~ 20 次抓取和抬升动作实验，每次实验的抓取力度都不相同，共得到 952 次有效抓取数据，其中滑动抓取和稳定抓取的比例接近 1∶1。在此基础上，参照文[11]所使用的数据增广方法，进行 9 倍增广，最终得到 8 568 组数据，每组数据包含 13 帧。数据集中 40 个物体的 6 795 组抓取数据作为训练集，剩余 10 个物体的 1 773 组抓取数据作为测试集。

图 5　数据集中的部分物体

Fig.5　Some objects in the dataset

3　实验结果与分析

为了测试所提出模型的性能，使用自主制作的数据集进行训练和测试。在 Pytorch 平台搭建模型，对网络参数随机初始化，使用交叉熵函数作为损失函数，Adam 优化器的学习率为 1×10^{-7}，batch 大小为 8。服务器平台使用 Intel Xeon E5-2620 CPU，48 GB 内存以及两块 GTX1080Ti GPU。接下来，从预训练网络、数据输入序列长度、输入模态和消融实验几方面进行测试与分析。

3.1　预训练网络

针对所提出的模型，使用预训练网络提取视觉数据的空间特征。为了比较不同的预训练网络在自主制作的数据集上的空间特征提取能力，分别使用 ResNet18、ResNet34、ResNet50、VGG-16 和 Inception-V3 网络进行测试。通过迁移学习的方法把这些网络在 ImageNet 数据集上预训练好的权重参数迁移到当前模型中。参考 ImageNet 中图片的大小，输入到网

络时将视觉图片大小调整为 224 × 224，输入序列长度预设为 13。对视触觉融合数据进行训练和测试的结果如表 2 所示。可以看出，不同的预训练网络即便使用同一数据集，对模型性能的影响也是很明显的。其中，使用 ResNet34 的性能最好。因此，在后续的测试中采用 ResNet34 提取视觉数据的空间特征。

表 2　不同预训练网络的测试结果

Tab.2　The testing results of different pre-training networks

预训练网络	准确率 /%
ResNet18	97.29
ResNet34	**98.98**
ResNet50	98.76
VGG-16	95.60
Inception-V3	96.39

3.2　数据输入序列长度

输入序列越长意味着有更多的信息输入到模型中，但更多的信息并不一定会使得检测效果更好。如果数据增加所带来的是更多的无用信息反而会使检测效果变差。使用 ResNet34，设定输入序列长度为 8 ~ 13，对视触觉融合数据进行训练和测试，结果如图 6 所示。可以看出，输入序列长度越长，准确率越高。当输入序列长度为 13 时准确率达到 98.98%，相比于输入序列长度为 8 时的准确率提升 3.15%。这说明对于本文的数据集，长序列所提供的有用信息比无用信息对模型的影响更大。从图 6 的曲线趋势可以预测，如果使用更长的输入序列，模型的检测准确率可能会达到更高，但考虑到数据集的大小、序列长度的限制等因素，在后续的实验中将择优选取输入序列长度为 13。

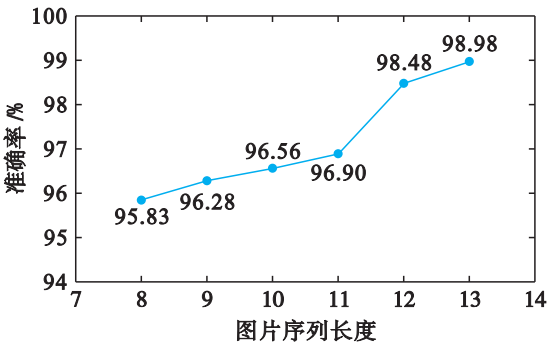

图 6　不同长度输入序列的精度曲线

Fig.6　The accuracy curve of input sequences of different lengths

3.3　数据输入模态

视觉模态和触觉模态具有各自不同的特点，其中视觉数据可提供物体的位置信息，触觉数据可提供与物体接触时的力触觉信息。为了验证视触觉融合能否提升模型的性能，进行了对比测试。单视觉模态输入时，使用 ResNet34 和 MS-TCN 提取数据的时空特征，并通过全连接层进行分类；单触觉模态输入时，使用如图 2 所示的触觉特征提取模型提取触觉数据的空间特征，通过 MS-TCN 提取数据的时序特征，最后经全连接层进行分类；视觉和触觉多模态输入时，使用所提出的视触觉融合模型进行特征提取和分类。测试结果如表 3 所示。可以看出，视触觉模态融合比单触觉模态或单视觉模态输入都有着更高的检测准确率。所提出的模型能够对视觉模态信息和触觉模态信息进行有效融合，对机器人抓取滑动的检测效果良好。

表 3　不同模态输入的测试结果

Tab.3　The testing results of different input modalities

输入模态	准确率 /%
单触觉	92.68
单视觉	96.00
视觉和触觉	**98.98**

3.4　消融实验

经过前面的综合对比测试确定，当输入数据为视觉模态和触觉模态、输入序列长度为 13、预训练网络为 ResNet34 时，模型能达到最佳效果。接下来，通过消融实验验证所提出模型性能的优越性。在消融实验中，分别使用 LSTM、TCN、MS-TCN 网络进行时序特征提取。从表 4 的实验结果可以看出，TCN 的性能比 LSTM 好。表明在滑动检测领域，相比于 LSTM，TCN 具有更为强健的时序特征提取能力。再者，使用 MS-TCN 进行时序特征提取得到的检测准确率相比 TCN 提升 0.45%。得益于 MS-TCN 的多分支结构，使其在每一层能够提取更加丰富的特征，这让其在 TCN 的基础上进一步提升时序特征提取能力和模型的检测准确率。最后，在使用 MS-TCN 的视触觉融合模型中，加入用于在视触觉特征拼接时重新分配权重的注意力机制，检测准确率提升 0.67%，高达 98.98%。由此验证，MS-TCN 和注意力机制的使用可有效提升所提出模型的性能。

在处理触觉数据时没有沿用常见的做法，即把触觉数据转换成触觉图片进行处理，而是直接对触觉原始数据进行特征提取。在此，进行如下对比实验。作为基准流程，首先参照文[13]对 XELA 触觉数据的处理方法，将 4 × 4 × 3 的触觉数据转换成

4 ×4 ×3 的 RGB 触觉图片，其中触觉数据的 X、Y、Z 通道分别对应触觉图片的R、G、B 通道。随后采用卷积神经网络提取触觉图片的空间特征，再通过 MS-TCN 提取其时序特征，最后经由全连接层分类输出。实验结果如表 5 所示，将 XELA 触觉数据转换为触觉图片再进行处理的方法得到的检测准确率仅为 75.51%，远低于直接使用触觉原始数据的触觉特征提取方法。另外，在进行视触觉融合时，使用触觉图片和视觉图片作为输入得到的检测准确率比使用触觉原始数据 + 视觉图片作为输入得到的检测准确率低 1.62%。通过实验可以证明所提出的触觉特征提取方法能够有效提取触觉数据的特征，并显著提升模型的性能。

表 4　消融实验结果

Tab.4　The results of ablation experiment

模型	准确率 /%
LSTM	97.01
TCN	97.86
MS-TCN	98.31
MS-TCN + Attention	**98.98**

表 5　触觉数据不同处理方法的对比实验结果

Tab.5　The results of different tactile data processing methods

输入模态	准确率 /%
触觉图片	75.51
触觉原始数据	**92.68**
触觉图片 + 视觉图片	96.69
触觉原始数据 + 视觉图片	**98.31**

3.5　泛化能力

为了进一步验证模型的泛化性能，使用文[11]中的数据集进行训练和测试，并与文[11]和文[12]提出的方法进行了对比。该数据集使用了 GelSight 触觉传感器，能够采集得到高分辨率的触觉图片，触觉图片大小为 640 ×480。为了能够提取触觉图片的空间特征，使用 ResNet34 预训练网路代替触觉特征提取网络。实验结果如表 6 所示。从结果可以看出，文[12]的方法无论是在本文的数据集上还是在文[11]的数据集上预测准确率都要比文[11]的方法高，这和文[12]的结论是一致的。此外，无论是使用阵列触觉传感器还是使用光学触觉传感器，所提出的模型均要比文[12]提出的方法有更高的预测准确率。这优秀的性能有助于将模型部署到使用不同类别触觉传感器的场景。

表 6　泛化性能实验

Tab.6　Generalization performance experiment

单位：%

方法	准确率	
	本文数据集	文[11]数据集
本文方法	**98.98**	**79.28**
文[12]方法	96.11	75.40
文[11]方法	93.69	73.09

4　总结

本文提出了一种结合注意力机制的视触觉融合模型，用于机器人在抓取物体时进行滑动检测。该模型首先使用 ResNet34 预训练网络提取视觉数据的空间特征，并通过所提出的触觉特征提取模型提取触觉数据的空间特征。使用 MS-TCN 网络提取视觉数据和触觉数据的时序特征。通过注意力机制在视触觉特征进行拼接时重新分配权重，并使用 MS-TCN 网络实现视触觉特征的信息融合。最后，经全连接层分类输出。此外，通过 Xarm 7DoF 机械臂、D455 RGB 摄像头和 XELA 触觉传感器对 50 个具有不同形状、材料和重量的日常物体执行抓取和抬升动作实验，采集相应的触觉数据和视觉数据，自主制作开放型数据集。基于该数据集对所提出模型进行训练和测试，在测试集上得到的检测准确率高达 98.98%。随后，通过消融实验和对比实验验证所提出的模型使用 MS-TCN 网络、CG 模块和触觉数据特征提取方法能够有效提升视触觉融合模型的性能。最后，经过泛化性能的实验证明了所提出的模型在不同类别的触觉传感器上均有优秀的表现。该工作有助于机器人在执行抓取类任务时进行高效的滑动检测，进而为及时调整抓取策略、确保任务的顺利执行奠定坚实的基础。

诚然，机器人灵巧抓取、视觉模态和触觉模态特征提取与融合仍然是挑战性问题，未来将会对更为高效的视触觉特征提取与融合方法、抓取策略等做深入的研究和尝试，从而对所提出模型的性能做进一步的优化与工程实践应用。

参考文献

[1] FENG Q, CHEN Z, DENG J, et al. Center-of-mass-based robust grasp planning for unknown objects using tactile-visual sensors [C]//2020 IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 610 – 617.

[2] SUN X, LIU T, ZHOU J, et al. Recent applications of different microstructure designs in high performance tactile sensors: A review[J]. IEEE Sensors Journal, 2021, 21(9): 10291 – 10303.
[3] LIU Y, BAO R, TAO J, et al. Recent progress in tactile sensors and their applications in intelligent systems[J]. Science Bulletin, 2020, 65(1): 70 – 88.
[4] CHEN W, KHAMIS H, BIRZNIEKS I, et al. Tactile sensors for friction estimation and incipient slip detection-Toward Dexterous robotic manipulation: A review[J]. IEEE Sensors Journal, 2018, 18(22): 9049 – 9064.
[5] JAMES J W, LEPORA N F. Slip detection for grasp stabilisation with a multi-fingered tactile robot hand[J]. IEEE Transactions on Robotics, 2021, 37(2): 506 – 519.
[6] GRIFFA P, SFERRAZZA C, D'ANDREA R. Leveraging distributed contact force measurements for slip detection: A physics-based approach enabled by a data-driven tactile sensor[C]//2022 IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2022: 4826 – 4832.
[7] SUI R, ZHANG L, LI T, et al. Incipient slip detection method for soft objects with vision-based tactile sensor[J/OL]. Measurement: Journal of the International Measurement Confederation, 2022, 203[2022 – 11 – 05]. https://www.sciencedirect.com/science/article/pii/S0263224122011022. DOI: 10.1016/j.measurement.2022.111906.
[8] GARCIA-GARCIA A, ZAPATA-IMPATA B S, ORTS-ESCOLANO S, et al. TactileGCN: A graph convolutional network for predicting grasp stability with tactile sensors[C/OL]//2019 IEEE International Joint Conference on Neural Networks. Piscataway, USA: IEEE, 2019[2022 – 11 – 21]. https://ieeexplore.ieee.org/document/8851984/. DOI: 10.1109/IJCNN.2019.8851984.
[9] JOHANSSON R S, VALLBO A B. Tactile sensibility in the human hand: Relative and absolute densities of four types of mechanoreceptive units in glabrous skin[J]. The Journal of Physiology, 1979, 286(1): 283 – 300.
[10] CALANDRA R, OWENS A, UPADHYAYA M, et al. The feeling of success: Does touch sensing help predict grasp outcomes?[C]//Conference on Robot Learning. Cambridge, MA: JMLR, 2017: 314 – 323.
[11] LI J, DONG S, ADELSON E. Slip detection with combined tactile and visual information[C]//2018 IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 7772 – 7777.
[12] YAN G, SCHMITZ A, P TOMO T, et al. Detection of slip from vision and touch[C]//2022 IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2022: 3537 – 3543.
[13] CUI S, WANG R, WEI J, et al. Grasp state assessment of deformable objects using visual-tactile fusion perception[C]//2020 IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 538 – 544.
[14] YAN G, SCHMITZ A, FUNABASHI S, et al. A robotic grasping state perception framework with multi-phase tactile information and ensemble learning[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6822 – 6829.
[15] ZAPATA-IMPATA B S, GIL P, TORRES F. Tactile-driven grasp stability and slip prediction[J/OL]. Robotics, 2019, 8(4)[2022 – 11 – 15]. https://www.mdpi.com/2218 – 6581/8/4/85. DOI: 10.3390/robotics8040085.
[16] LEA C, FLYNN M D, VIDAL R, et al. Temporal convolutional networks for action segmentation and detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 1003 – 1012.
[17] BAI S, KOLTER J, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[EB/OL]. (2018 – 04 – 19)[2022 – 11 – 01]. https://arxiv.org/abs/1803.01271.
[18] MARTINEZ B, MA P, PETRIDIS S, et al. Lipreading using temporal convolutional networks[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, USA: IEEE, 2020: 6319 – 6323.
[19] CHUNG J S, SENIOR A, VINYALS O, et al. Lip Reading sentences in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 3444 – 3453.
[20] YANG S, ZHANG Y H, FENG D, et al. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild[C/OL]//2019 IEEE International Conference on Automatic Face & Gesture Recognition. Piscataway, USA: IEEE, 2019[2022 – 10 – 28]. https://ieeexplore.ieee.org/document/8756582/. DOI: 10.1109/FG.2019.8756582.
[21] NIU Z, ZHONG G, YU H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48 – 62.
[22] TU Z, LIU Y, LU Z, et al. Context gates for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 87 – 99.

作者简介

黄兆基(1997 –), 男, 硕士生。研究领域为机器人抓取, 滑动检测, 多模态融合。
高军礼(1973 –), 男, 博士, 副教授。研究领域为机器人感知与操作控制, 机器人人机协作与交互。
唐兆年(1997 –), 男, 硕士生。研究领域为机器人灵巧操作, 机器人抓取构型。