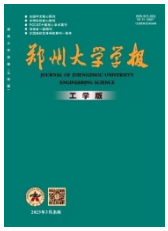

郑州大学学报(工学版)
Journal of Zhengzhou University(Engineering Science)
ISSN 1671-6833,CN 41-1339/T

《郑州大学学报(工学版)》网络首发论文

题目：多模态数据融合的加工作业动态手势识别方法
作者：张富强，曾夏，白筠妍，丁凯
DOI：10.13705/j.issn.1671-6833.2024.02.007
收稿日期：2023-09-01
网络首发日期：2023-12-04
引用格式：张富强，曾夏，白筠妍，丁凯．多模态数据融合的加工作业动态手势识别方法[J/OL]．郑州大学学报(工学版).
https://doi.org/10.13705/j.issn.1671-6833.2024.02.007

网络首发时间：2023-12-04 12:13:55
网络首发地址：https://link.cnki.net/urlid/41.1339.T.20231201.1726.007

多模态数据融合的加工作业动态手势识别方法

张富强[1,2]，曾　夏[1,2]，白筠妍[1,2]，丁　凯[1,2]

（1.长安大学 道路施工技术与装备教育部重点实验室，陕西 西安 710064；2.长安大学 智能制造系统研究所，陕西 西安 710064）

摘　要：为了解决单模态数据所提供的特征信息缺乏而导致的识别准确率难以提高、模型鲁棒性较低等问题，提出了面向人机交互的加工作业多模态数据融合动态手势识别策略。首先，采用 C3D 网络模型基于视频的空间维度和时间维度对深度图像和彩色图像 2 种模态数据进行特征提取；其次，将 2 种模态数据识别结果在决策层按最大值规则进行融合，同时，将原模型使用的 Relu 激活函数替换为 Mish 激活函数优化梯度特性；最后，通过 3 组对比实验得到 6 种动态手势的平均识别准确率达到 96.8%，相比彩色和深度单模态数据，分别提升了 2.98%和 2.76%。结果表明：所提方法实现了加工作业中动态手势识别的高准确率和高鲁棒性的目标，对人机交互技术在实际生产场景中的应用起到推动作用。

关键词：多模态数据融合；加工作业；动态手势识别；C3D；Mish 激活函数；人机交互

中图分类号：TH166　　**文献标志码**：A　　**doi**：10.13705/j.issn.1671-6833.2024.02.007

智能制造在工业场景中的渐次落地对复杂加工装备的数字化、网络化和智能化演变提出要求，人机交互技术赋能复杂加工装备的智能化升级，成为当前的研究热点问题[1]。手势在交互过程中为非接触方式，并且作为人的本能身体语言，具有简单易学、语义丰富等优点，而动态手势包含手势的运动轨迹和形态信息，进一步提高了动作区分和识别的准确性。采用动态手势作为人机交互的输入，对复杂加工装备进行操作与交互，可为车间加工生产实现效率最大化提供帮助。

基于视觉的传统手势识别通常由手部检测分割、手部跟踪、手部图像信息提取及识别这几个步骤分步展开。识别阶段最为广泛使用的方法有动态时间规整（DTW）算法[2]、隐马尔可夫（HMM）模型[3]以及支持向量机（SVM）等。李浩等[4]采用动态时间规整和支持向量机算法进行动态手势识别，取得了优异的可靠性和稳定性。Jia 等[5]提出一种基于手骨骼信息和隐马尔可夫模型的非轨迹手势识别算法。彭金柱等[6]结合了视觉和肌电信息进行时域特征融合，采用支持向量机进行模型训练，有效提高了识别准确率。但上述方法在表达能力、鲁棒性、灵活性及可拓展性方面存在不足，难以满足复杂场景中的实际需要。

基于深度学习的手势识别方法是通过深度神经网络对原始图像和视频自动学习提取特征以及完成分类，具有较强的泛化能力和适应性[7]。其中基于卷积神经网络（CNN）、三维卷积神经网络（3D-CNN）以及序列模型的方法得到了广泛的应用探索。Cheok 等[8]对手势和手语识别中的人工智能算法进行了综述。Mujahid 等[9]提出一种基于 YOLOv3 和 DarkNet-53 卷积神经网络的手势识别模型。Adithya 等[10]针对复杂背景下的手势识别问题，提出了一种基于深度卷积神经网络的模型结构，准确率达到 90%以上。Sharma 等[11]设计了一种结构更紧凑的卷积神经网络模型，能在训练时间较短的情况下取得较好的识别效果。刘杰等[12]设计了一种基于卷积视觉自注意力模型的多尺度时空特征融合网络，在 Jester 动态手势数据集上识别率达到 92.26%。Singh 等[13]和 Zhang 等[14]利用 3D-CNN 实现了高精度动态手势识别，较 2D-CNN 能更好捕

收稿日期：2023-09-01；**修订日期**：2023-11-16
基金项目：国家重点研发计划项目（2021YFB3301702）；陕西省科技重大专项（2018zdzx01-01-01）
作者简介：张富强（1984—），男，山西运城人，长安大学副教授，博士，主要从事面向人机交互的智能制造方面的研究，E-mail：fqzhang@chd.edu.cn。

获时间维度上的判别特征。Yang 等[15]提出一种双层双向递归神经网络,用于从 Leap Motion 控制器识别动态手势,表现出良好的性能。Khodabandelou 等[16]使用基于注意力的递归神经网络来捕获手部运动的时间特征。Li 等[17]提出一种双流神经网络,以基于自注意的图卷积网络(SAGCN)提取短期时间信息和层次空间信息,使用残差连接增强双向独立循环神经网络(IndRNN)提取长期时态信息。CNN(或 3D-CNN)与序列模型相结合的方式也表现出不错的性能。Palanisamy 等[18]和谷学静等[19]结合深度卷积神经网络与长短期记忆网络,提高了识别精度和抗干扰能力。Rehman 等[20]利用 3D-CNN 提取光谱和空间特征,将其提供给 LSTM 网络进行分类,通过对比实验证明了其优越性。综上所述,随着神经网络和深度学习的不断深入发展,手势识别算法不断优化,使得目标识别准确率不断提高。但实际工程应用中复杂的光照条件、背景干扰及个体间的差异导致单模态数据所提供的特征信息难以支撑模型达到所需的准确性和鲁棒性。

针对上述问题,本文提出一种面向人机交互的加工作业多模态数据融合动态手势识别策略,利用 Kinect 传感器采集动态手势数据集,采用 C3D 网络模型对深度图像和彩色图像 2 种模态数据进行特征提取并进行决策层融合,同时对激活函数进行优化改善“神经元坏死”问题,能有效提高识别准确率,通过 Unity 3D 实现手势交互,验证了所提方法的有效性。

1 问题描述

在基于复杂加工装备的人机交互过程中,以手势作为交互的方式具有丰富的内涵[21]。将复杂加工装备运行数据映射为孪生模型,供远程操作人员进行操控,以此实现车间内部的场景共享。在此基础上,可以利用与服务器相连接的 Kinect 相机采集操作人员的不同手势输入,并经过多模态数据融合的加工作业动态手势识别技术对输入手势进行处理,实现操作人员远程手势控制复杂加工装备的运行。在实际加工作业中,可以优化生产流程,使生产过程更加高效和可控,降低生产成本。具体实现过程如图 1 所示。

2 基于 3D-CNN 的动态手势识别

对一组输入图像数据,用 b 表示 batch size 的大小,c 表示颜色通道数量,n 表示进行堆叠的视频中图像帧的数量,h 和 w 分别表示视频图像帧的高和

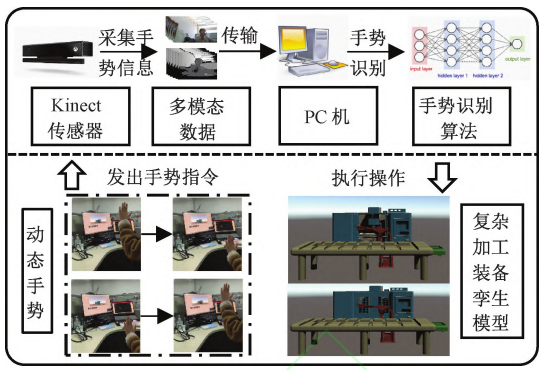

图 1 动态手势交互示意图

Figure 1 Dynamic gesture interaction diagram figure

宽。三维卷积过程的计算公式如式(1)所示[22]:

$$v_{ij}^{xyz}=f\left(\sum_{d}\sum_{r=0}^{R_{i-1}}\sum_{s=0}^{S_{i-1}}\sum_{t=0}^{T_{i-1}}W_{ijd}^{rst}\cdot P_{(i-1)d}^{(x+r)(y+s)(r+t)}+b_{ij}\right)。\quad(1)$$

式中:P 表示卷积过程的输入;x 和 y 为空间维度;z 表示时间维度;$f(\cdot)$ 表示激活函数;v_{ij}^{xyz} 表示第 i 层的 j 特征图在 (x,y) 位置的输出;W_{ijd}^{rst} 表示在当前层卷积核的权重;b_{ij} 表示第 i 层的 j 卷积核的特征偏置量;R_i、S_i、T_i 分别表示深度、高、宽;参数 r、s、t 表示本次卷积的值。

2.1 C3D 模型

Tran 等[23]在 3D-CNN 网络模型的基础上进行优化,提出 C3D 的网络模型,该模型能够同时获取外观信息和运动信息,具有简洁紧凑的网络结构,并在视频分类和动作识别等任务中,具有较好的性能表现,取得较高的识别准确率。该网络结构如图 2 所示。

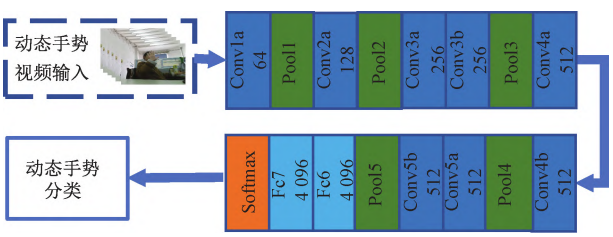

图 2 C3D 的网络结构

Figure 2 Network structure of C3D

通过图 2 可以看出,C3D 网络主要包括 3D 卷积层、3D 池化层和全连接层,激活函数为 Relu 函数,将提取的特征进行非线性处理,提高网络模型的表达能力。Conv 表示 3D 卷积操作,Pool 表示 3D 池化操作。C3D 网络模型的详细参数如表 1 所示。

2.2 激活函数改进

C3D 初始模型采用了 Relu 函数作为激活函数,然而,Relu 函数的负抑制特性存在一些不足之处,因为在反向传播过程中,当输入为负时,梯度会完全为零,神经元无法继续更新,可能会导致大量神经元

出现“坏死”。

表 1　C3D 网络模型参数表

Table 1　C3D network model parameter table

层名	输出大小	卷积核	步长	填充
Conv1a	64	(3,3,3)	(1,1,1)	(1,1,1)
Pool1	64	(1,2,2)	(1,2,2)	(0,0,0)
Conv2a	128	(3,3,3)	(1,1,1)	(1,1,1)
Pool2	228	(2,2,2)	(2,2,2)	(0,0,0)
Conv3a	256	(3,3,3)	(1,1,1)	(1,1,1)
Conv3b	256	(3,3,3)	(1,1,1)	(1,1,1)
Pool3	256	(2,2,2)	(2,2,2)	(0,0,0)
Conv4a	512	(3,3,3)	(1,1,1)	(1,1,1)
Conv4b	512	(3,3,3)	(1,1,1)	(1,1,1)
Pool4	512	(2,2,2)	(2,2,2)	(0,0,0)
Conv5a	512	(3,3,3)	(1,1,1)	(1,1,1)
Conv5b	512	(3,3,3)	(1,1,1)	(1,1,1)
Pool5	512	(2,2,2)	(2,2,2)	(0,1,1)
Fc6	4 096			
Fc7	4 096			
Softmax	6			

为了有效避免上述问题,采用 Mish 激活函数代替 Relu 激活函数,Mish 激活函数[24]的数学表达式如下:

$$f(x) = x\tanh(\ln(1 + e^x))。\tag{2}$$

式中:x 表示输入值;$f(x)$ 表示输出值。

Mish 激活函数曲线如图 3 所示,Mish 激活函数具有以下优点:

(1)Mish 激活函数的梯度更加平滑,具有良好的梯度下降效果。

(2)Mish 激活函数没有边界限制,可以有效避免饱和问题。

(3)在负半轴处,允许流入较小梯度,避免神经元“坏死”的负抑制现象。

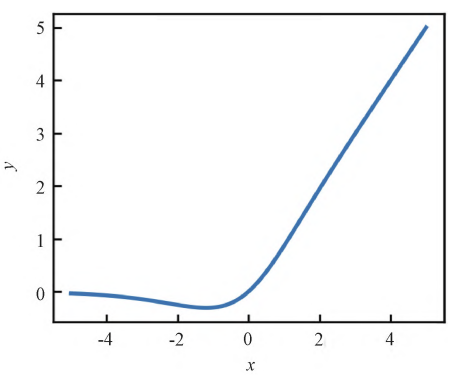

图 3　Mish 激活函数

Figure 3　Mish activation function

3　多模态数据融合的动态手势识别模型

单模态数据手势识别一般情况下使用单个数据模态(通常为 RGB 图像)进行训练和测试,而多模态数据手势识别采用不同的数据模态进行训练,并在测试期间对多模态数据预测结果进行分类。多模态数据融合的动态手势识别采用 C3D 模型对多种模态的手势数据训练学习,即利用多模态数据来训练网络模型得到不同性能的网络模型后对动态手势进行测试。多模态数据融合方法通常包括数据层融合、特征层融合和决策层融合[25]。

根据不同模态数据的特点,选用深度手势数据和彩色手势数据 2 种模态数据作为输入。考虑到数据层融合和特征层融合存在不同模态间互相干扰以及模型复杂度等问题,决策层融合对外界影响因素干扰具有很好的鲁棒性,充分考虑不同模态特征的差异性,因此使用决策层融合方法进行融合。此种融合策略可以减少模型的计算量,提高运行速度与模型整体的识别准确率。将彩色手势视频和深度手势视频分别输入到 C3D 网络模型中进行训练识别,然后将各自的识别结果在决策层进行融合后得到最终的结果。

所采用的 C3D 模型包含 8 个卷积层、5 个池化层、2 个全连接层以及一个 Softmax 输出层。卷积层的卷积核均为 3×3×3,步长为 1×1×1,每个卷积层均采用 Mish 激活函数。除第一个池化层的核与步长设置为 2×2×1 外,其余池化层的核及步长均设置为 2×2×2。全连接层输出单元为 4 096,在其后使用 dropout,防止出现过拟合现象。Softmax 层的输出设置为 6 个单元,对应 6 种手势类别。多模态数据融合识别模型结构如图 4 所示。

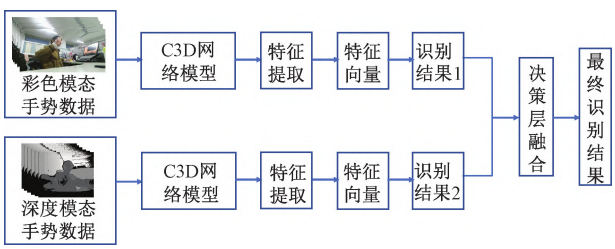

图 4　多模态数据融合识别模型结构

Figure 4　Multi-modal data fusion identification model structure

采用交叉熵损失函数作为模型的损失函数:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log p_{ik}。\tag{3}$$

式中:$Loss$ 为模型损失;i 代表样本序号;N 为该轮训练的样本总数;K 为类别总数;若 i 样本预测为 k 类别,则 y_{ik} 值为 1,否则为 0;p_{ik} 代表第 i 个样本由 Softmax 层输出的第 k 个类别的概率。多模态融合动态手势识别算法伪代码如下。

1. 输入参数:迭代次数 *nEpochs*、批样本数 *batch_size* 等。
2. 初始化 C3D 模型参数 θ_{RGB} 和 θ_{depth}。
3. C3D 模型训练。
3.1. for $i = 1$:*nEpochs*
从彩色数据集中采样,使用式(3)更新 θ_{RGB};
End for
3.2. for $i = 1$:*nEpochs*
从深度数据集中采样,使用式(3)更新 θ_{depth};
End for
4. 多模态数据决策层融合。
4.1. 彩色数据输入 C3D 模型得到预测特征向量 $\boldsymbol{X}_1$。
4.2. 深度数据输入 C3D 模型得到预测特征向量 $\boldsymbol{X}_2$。
4.3. 将特征向量 $\boldsymbol{X}_1$ 与 $\boldsymbol{X}_2$ 按最大值规则进行特征层融合,输出最终的手势分类结果。
5. end。
6. Return 最终的手势分类结果。

4 案例验证

4.1 动态手势数据集的采集

动态手势数据集是由实验室采用 Kinect 传感器的彩色摄像头和深度摄像头同时进行拍摄的自制数据集。将 Kinect 连接到 PC 计算机上,并下载安装 Kinect for windows SDK 2.0、OpenCV,在 Visual Studio 2019 下完成 Kinect 和 OpenCV 的环境配置,使用 C#语言进行程序编写,设置视频帧数,实现 2 种模态手势数据的采集以及自动保存。

选取 6 种动态手势作为代表手势,包括 2 种双手手势和 4 种单手手势。双手手势为双手由里向外移动(inside_to_outside)和双手由外向里移动(outside_to_inside),单手手势为左手从右向左移动(right_to_left)、左手从左往右移动(left_to_right)、左手从下向上移动(down_to_up)、左手从上向下移动(up_to_down)。从以下 4 个方面规范化自制数据集。

(1)样本类别和数量:采用 Kinect 的彩色摄像头和深度摄像头对 6 种手势进行不同角度、不同背景、不同光照的手势数据集采集。每种手势采集 100 组,一共得到 600 组手势数据集。

(2)标签和注释:对每个手势样本给出明确的标签或注释。

(3)数据格式和存储:使用统一的视频文件.avi 数据格式,同时该数据集含有基于彩色图的图像信息和基于深度图的深度信息 2 种模态,均以 15 帧/s 的速度采集,手势样本长度为 50~70 帧,从而在一定程度上使不同速度的手势识别更准确。

(4)多样性和多尺度特性:从多个角度多种距离采集 5 名人员的手势,允许手指的稍微变形,充分保证数据集的多样性与多尺度特性。

4.2 数据集划分与参数设置

(1)实验设置了训练集、验证集和测试集。训练集是用于训练模型的数据集,训练好的模型经验证集进行调整优化,最后由测试集评估模型的性能和泛化能力。在实验准备阶段,对视频数据集进行数据划分,划分比例如图 5 所示。其中,训练集为 384 个视频,验证集为 96 个视频,测试集为 120 个视频。在动态手势识别过程中,首先需要将动态视频转换为图像帧,从样本视频中每隔 4 帧截取出一张图像数据,一共截取 16 帧长的帧序列,如果不够 16 帧则减少间隔所需帧数。此外,将每一帧图片规范为 112×112 大小,形成统一的输入以便进行后续识别。

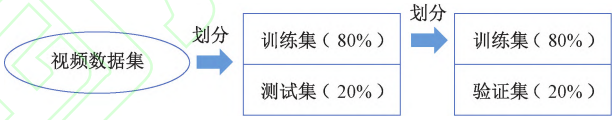

图 5 视频数据集的划分

Figure 5 Partitioning of video datasets

(2)参数设置。在训练网络模型中采用随机梯度下降算法(SGD)进行模型优化,迭代次数设为 50;批样本数设为 4;权重衰减取 0.000 05;动量衰减取 0.9;学习率设置为 0.000 01。

4.3 实验结果与对比分析

(1)多模态数据融合的实验分析。实验采用多模态数据融合的 C3D 网络进行动态手势识别。图 6(a)和图 6(b)分别展示了深度视频数据和彩色视频数据在训练集和验证集的准确率变化过程。由图 6 可以看出彩色视频数据和深度视频数据在训练集和验证集中,由于有着各自的图像特点,准确率高低交错。多模态数据融合即对 2 种结果进行融合,同一时刻根据最大值规则,选取准确率最大值作为最终识别结果,因此,本文提出在决策层融合的方法可以保证最佳准确率,提升稳定性。

为了进一步验证融合策略在测试集上的表现,分别统计 6 种动态手势在不同模态下的准确率和融合后的准确率对比,如表 2 所示。由表 2 可以看出,相比单模态数据,采用 2 种模态数据融合后的 6 种动态手势准确率均得到提升,融合后的平均准确率达到 96.8%。

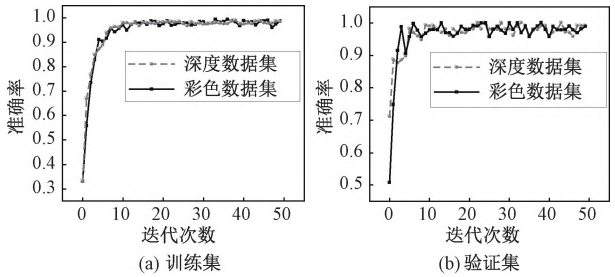

图 6　不同模态数据集的准确率变化曲线

Figure 6　Accuracy curve of different modal datasets

表 2　测试视频识别精准率

Table 2　Recognition accuracy of the test videos　%

手势类别	彩色模态的准确率	深度模态的准确率	融合后的准确率
up_to_down	94.0	93.6	97.9
down_to_up	95.6	94.3	96.0
inside_to_outside	93.7	95.0	96.8
outside_to_inside	92.8	94.9	98.4
left_to_right	93.6	92.4	95.6
right_to_left	94.4	95.0	96.2

(2)改进激活函数的对比实验分析。为了验证改进激活函数的作用,将 Relu 函数改为 Mish 激活函数后进行 2 组对比实验,分别在训练集和验证集上对比准确率和损失函数的走向趋势,结果如图 7 和图 8 所示。图 7(a)和图 7(b)分别为不同激活函数在训练集和验证集上准确率的变化曲线对比图;图 8(a)和图 8(b)为损失函数的变化曲线对比图。

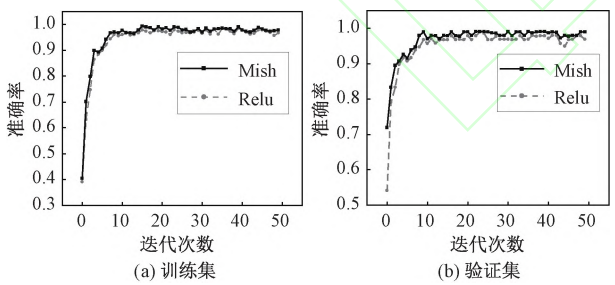

图 7　改进后激活函数准确率对比

Figure 7　Comparison of accuracy of activation function after improvement

通过图 7 可以看出,采用 Mish 激活函数的识别准确率高于采用 Relu 激活函数的识别准确率;通过图 8 可知,采用 Mish 激活函数的 C3D 网络模型与采用 Relu 激活函数的 C3D 网络模型相比,其损失函数值更小且收敛效果更加显著。说明改进后的激活函数的网络模型性能优于之前的网络模型性能。因为 Mish 激活函数允许较小的负梯度流入从而避免部分神经元“坏死”,同时在梯度下降过程中可以使网络更好地找到相对最佳点。

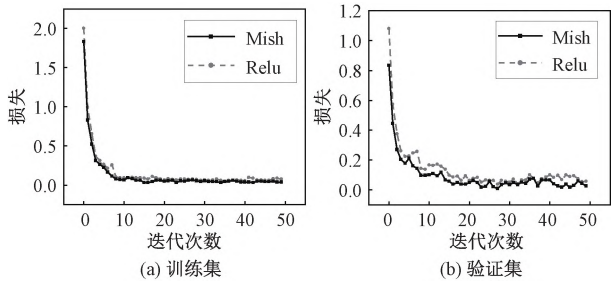

图 8　改进后激活函数的损失函数对比

Figure 8　The loss function comparison diagram of the improved activation function

4.4　应用案例

通过 Kinect v2.9.unitypackage 插件联合 Kinect 2.0 和 Unity 3D 虚拟软件进行原型系统的开发,在 Unity 3D 中建立孪生模型如图 9 所示。

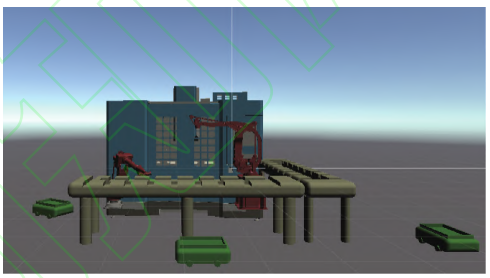

图 9　复杂加工装备孪生体示意图

Figure 9　Complex processing equipment digital twin

本次案例以 2 种动态手势为例,验证了将所提方法用于复杂加工装备孪生模型的交互控制的可行性。指定双手控制,右手不动,左手向上方移动时对应防护门开启指令;右手不动,左手向下方移动时对应防护门关闭指令。

操作人员根据自己的手部动作与 Unity 里的复杂加工装备孪生体进行交互。当操作人员根据需求做出手势动作时,Unity 里的模型会遵循定义好的手势指令给出相应反馈。

当在 Kinect 传感器前做左手从下往上的动作,可以在 Unity 界面中看见防护门开启,当在 Kinect 传感器前做左手从上往下的动作,可以在 Unity 界面中看见防护门关闭,如图 10 所示。

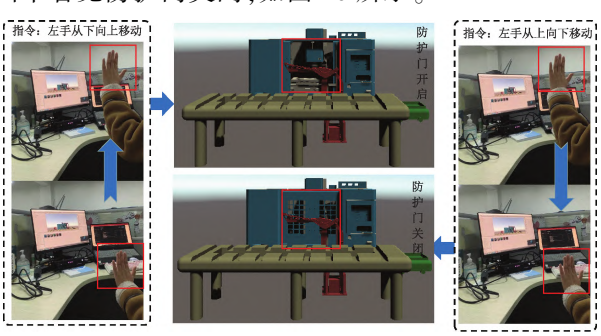

图 10　动态手势交互结果

Figure 10　Dynamic gesture interaction results

为了进一步验证案例可行性,对这 2 种手势进行测试,每种手势测试 100 次,结果如表 3 所示。

表 3　不同手势测试的实验结果

Table 3　Experimental results of different gesture tests

手势名称	测试次数	交互正确次数	准确率/%
down_to_up	100	95	95
up_to_down	100	97	97

通过表 3 可以看出,动态手势对复杂加工装备孪生体的交互有着良好的交互结果,正确识别且交互结果正确的准确率为 96%,表明本文提出的方法在与孪生模型交互中的可行性和有效性。

5　结论

(1)研究了基于多模态数据融合的加工作业动态手势识别,提出了三维卷积神经网络和多模态数据融合的方法。

(2)对改进后的激活函数进行描述,然后阐述了使用的多模态数据融合的动态手势识别模型。在实验部分,采用 Kinect 传感器同时采集深度和彩色 2 种模态的动态手势数据集,选用 Mish 激活函数,通过 C3D 模型进行训练并在决策层进行融合得到识别结果。

(3)实验结果表明,融合后的平均准确率达到 96.8%,且通过对比实验证明对激活函数进行改进能提高模型的准确率、改善收敛性,验证了所提方法的合理性和有效性,最后联合 Kinect 和 Unity 平台开发原型系统证明了所提方法在与孪生模型交互中的可行性。

参考文献:

[1] 李浩, 刘根, 文笑雨, 等. 面向人机交互的数字孪生系统工业安全控制体系与关键技术[J]. 计算机集成制造系统, 2021, 27(2): 374-389.
LI H, LIU G, WEN X Y, et al. Industrial safety control system and key technologies of digital twin system oriented to human-machine interaction[J]. Computer Integrated Manufacturing Systems, 2021, 27(2): 374-389.

[2] LIU N J, LOVELL B C, KOOTSOOKOS P J, et al. Model structure selection & training algorithms for an HMM gesture recognition system[C]//Ninth International Workshop on Frontiers in Handwriting Recognition. Piscataway: IEEE, 2004: 100-105.

[3] HARTMANN B, LINK N. Gesture recognition with inertial sensors and optimized DTW prototypes[C]//2010 IEEE International Conference on Systems, Man and Cybernetics. Piscataway: IEEE, 2010: 2102-2109.

[4] 李浩, 杨森林, 张晓丽. 基于机器视觉的火车驾驶员动态手势识别方法[J]. 传感器与微系统, 2021, 40(2): 34-37, 43.
LI H, YANG S L, ZHANG X L. Dynamic gesture recognition method of train driver based on machine vision[J]. Transducer and Microsystem Technologies, 2021, 40(2): 34-37, 43.

[5] JIA L S, ZHOU X Z, XUE C Q. Non-trajectory-based gesture recognition in human-computer interaction based on hand skeleton data[J]. Multimedia Tools and Applications, 2022, 81(15): 20509-20539.

[6] 彭金柱, 董梦超, 杨扬. 基于视觉和肌电信息融合的手势识别方法[J]. 郑州大学学报(工学版), 2021, 42(2): 67-73.
PENG J Z, DONG M C, YANG Y. Human gesture recognition method based on vision and EMG signal information[J]. Journal of Zhengzhou University (Engineering Science), 2021, 42(2): 67-73.

[7] CHAKRAVARTHI S S, RAO B N K, CHALLA A P, et al. Gesture recognition for enhancing human computer interaction[J]. Journal of Scientific & Industrial Research, 2023, 82(4): 438-443.

[8] CHEOK M J, OMAR Z, JAWARD M H. A review of hand gesture and sign language recognition techniques[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(1): 131-153.

[9] MUJAHID A, AWAN M J, YASIN A, et al. Real-time hand gesture recognition based on deep learning YOLOv3 model[J]. Applied Sciences, 2021, 11(9): 4164.

[10] ADITHYA A, RAJESH R. A deep convolutional neural network approach for static hand gesture recognition[J]. Procedia Computer Science, 2020, 171: 2353-2361.

[11] SHARMA S, SINGH S. Vision-based hand gesture recognition using deep learning for the interpretation of sign language[J]. Expert Systems with Applications, 2021, 182: 115657.

[12] 刘杰, 王月, 田明. 多尺度时空特征融合的动态手势识别网络[J]. 电子与信息学报, 2023, 45(7): 2614-2622.
LIU J, WANG Y, TIAN M. Dynamic gesture recognition network based on multiscale spatiotemporal feature fusion[J]. Journal of Electronics & Information Technology, 2023, 45(7): 2614-2622.

[13] SINGH D K. 3D-CNN based dynamic gesture recognition for Indian sign language modeling[J]. Procedia Computer Science, 2021, 189: 76-83.

[14] ZHANG W J, WANG J C. Dynamic hand gesture recognition based on 3D convolutional neural network models[C]//2019 IEEE 16th International Conference on Net-

working, Sensing and Control (ICNSC). Piscataway: IEEE, 2019: 224-229.

[15] YANG L C, CHEN J, ZHU W H. Dynamic hand gesture recognition based on a leap motion controller and two-layer bidirectional recurrent neural network[J]. Sensors, 2020, 20(7): 2106.

[16] KHODABANDELOU G, JUNG P G, AMIRAT Y, et al. Attention-based gated recurrent unit for gesture recognition[J]. IEEE Transactions on Automation Science and Engineering, 2021, 18(2): 495-507.

[17] LI C K, LI S, GAO Y B, et al. A two-stream neural network for pose-based hand gesture recognition[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(4): 1594-1603.

[18] PALANISAMY G, THANGASWAMY S S. An efficient hand gesture recognition based on optimal deep embedded hybrid convolutional neural network-long short term memory network model[J]. Concurrency and Computation: Practice and Experience, 2022, 34(21): 1-15.

[19] 谷学静, 周自朋, 郭宇承, 等. 基于 CNN-LSTM 混合模型的动态手势识别方法[J]. 计算机应用与软件, 2021, 38(11): 205-209.
GU X J, ZHOU Z P, GUO Y C, et al. Dynamic gesture recognition method based on cnn-lstm hybrid model[J]. Computer Applications and Software, 2021, 38(11): 205-209.

[20] REHMAN M, AHMED F, KHAN M, et al. Dynamic hand gesture recognition using 3D-CNN and LSTM networks[J]. Computers Materials & Continua, 2021, 70(3): 4675-4690.

[21] MO D H, TIEN C L, YEH Y L, et al. Design of digital-twin human-machine interface sensor with intelligent finger gesture recognition[J]. Sensors, 2023, 23(7): 3509.

[22] JI S W, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[23] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2015: 4489-4497.

[24] MERCIONI M A, HOLBAN S. Soft clipping mish-A novel activation function for deep learning[C]//2021 4th International Conference on Information and Computer Technologies (ICICT). Piscataway: IEEE, 2021: 13-17.

[25] 杨婷. 基于深度学习的多模态情感识别[D]. 南昌: 南昌大学, 2021.
YANG T. Multimodal emotion recognition based on deep learning[D]. Nanchang: Nanchang University, 2021.

Dynamic Gesture Recognition Method for Machining Operations Based on Multi-modal Data Fusion

ZHANG Fuqiang[1,2], ZENG Xia[1,2], BAI Junyan[1,2], DING Kai[1,2]

(1. Key Laboratory of Road Construction Technology and Equipment of MOE, Chang'an University, Xi'an 710064, China; 2. Institute of Smart Manufacturing Systems, Chang'an University, Xi'an 710064, China)

Abstract: In order to solve the problems that the lack of feature information provided by single mode data makes it difficult to improve the recognition accuracy and the low robustness of the model, a dynamic gesture recognition strategy based on multi-modal data fusion of machining operations for human-computer interaction was proposed. Firstly, the C3D network model was used to extract features from the depth image and color image modal data based on the spatial and temporal dimensions of videos. Secondly, the recognition results of the two modal data were fused according to the maximum principle at the decision-making level. Meanwhile, the Relu activation function used in the original model was replaced by Mish activation function to optimize the gradient update effect. Finally, through three sets of comparative experiments, it was found that the average recognition accuracy of six dynamic gestures reached 96.8%, which is 2.98% and 2.76% higher than that of color and depth single-mode data respectively. The results show that the proposed method achieves the goal of high accuracy and high robustness of dynamic gesture recognition in machining operation, which plays a role in promoting the application of human-computer interaction technology in actual production scenes.

Keywords: multi-modal data fusion; machining operation; dynamic gesture recognition; C3D; Mish activation function; human-computer interaction