邵晨悦, 孟青云, 查佳佳, 等. 基于视觉识别技术的手势自动跟随研究[J]. 智能计算机与应用,2024,14(11):117-123.
DOI:10.20169/j.issn.2095-2163.241118

基于视觉识别技术的手势自动跟随研究

邵晨悦[1,2], 孟青云[1], 查佳佳[2], 熊亦可[1], 严加勇[1]
(1 上海健康医学院 医疗器械学院, 上海 201318; 2 上海理工大学 健康科学与工程学院, 上海 200093)

摘 要: 为了构建一个跨平台的实时手势跟踪应用系统,本文首先利用卷积神经网络从单色图像中估计手部的姿势和形状,并与直接在图像中预测的二维坐标和三维坐标结果进行比较实验证明。实验结果表明,预测的二维坐标与三维坐标的差异不大,可以直接进行三维坐标的预测。其次,基于跟踪到的 21 个手关节的坐标位置,进一步提出了计算手指关节运动角度的方法,并在舵机上进行角度转动复现实验证明。同时,实验证明在计算关节角度方面,所得到的角度与实际角度之间的误差控制在 10°以内,并且该方法能够准确实时地反映手部动作。最终,本研究搭建了一个结合识别、检测、分类和跟踪多功能的跨平台实时手势自动跟随系统,可以在单根手指上实现多种功能。

关键词: 视觉识别; 卷积神经网络; MediaPipe; 手势跟随; 单片机电机控制
中图分类号: TP391.41 **文献标志码**: A **文章编号**: 2095-2163(2024)11-0117-07

Research on automatic gestures following based on visual recognition technology

SHAO Chenyue[1,2], MENG Qingyun[1], ZHA Jiajia[2], XIONG Yike[1], YAN Jiayong[1]
(1 School of Medical Instrumentation, Shanghai University of Medicine and Health Sciences, Shanghai 201318, China;
2 School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In order to construct a cross platform real-time gesture tracking application system, this paper firstly utilizes convolutional neural networks to estimate hand poses and shapes from monochrome images, and compares the results of directly predicting 2D and 3D coordinates in the image through experimental verification. The experimental results show that there is little difference between the predicted 2D coordinates and the 3D coordinates, indicating that direct prediction of 3D coordinates is feasible. Furthermore, based on the tracked coordinates of 21 hand joints, a method for calculating finger joint motion angles is proposed and verified through experiments involving angle rotation replication on servos. The experiments demonstrate that the calculated joint angles have an error within 10°compared to the actual angles, and this method can accurately and in real-time reflect hand movements. Finally, this research establishes a cross-platform real-time gesture-based automatic tracking system integrated with recognition, detection, classification, and tracking functionalities, capable of achieving multiple functions on a single finger.
Key words: visual recognition; convolutional neural network; MediaPipe; gestures tracking; microcontroller motor control

0 引 言

手势跟随作为一种基于视觉识别技术的关键输入信息,在现代人机交互体系中发挥着重要作用。目前,市场上的静态手势识别已经成为一种图像化输入指令,可以执行相应的计算机操作。

在动态手势识别方面,国内外的研究团队都展开了大量的研究。例如,已有学者提出了基于肌电信号[1]、超声波信号[2]和图像[3]的识别方法。这些方法都有各自的优点和适用范围,但大多数方法都是用于分类问题,并没有深入探讨在跟踪方面的应用。此外,在研究中选择适当的识别方法和特征处理也变得至关重要。为了提高动态手势识别的准确性和适用范围,自动跟踪技术得以迅速发展。

在手势跟踪方面,国内外的研究者们致力于提高手势跟踪系统的连续性和实时性。例如,詹金峰[4]基于网络摄像头获取的视频数据,提高了对同义背景干扰的抵抗能力,实现了高准确率的手势跟

作者简介: 邵晨悦(2001—),女,硕士研究生,主要研究方向:生物医学工程。
通信作者: 孟青云(1971—),女,博士,高级工程师,主要研究方向:柔性医疗器械设计及智能控制,康复机器人设计及智能控制,机器视觉等。
Email: mengqy@sumhs.edu.cn。

收稿日期: 2023-06-19

踪并满足实时性要求。张继凯等学者[5]讨论了图像特征对三维手势跟踪的影响,并总结了缓解图像深度模糊的方法。Xiao 等学者[6]在 MediaPipe 框架下测试了视觉、听觉和表面肌电信号在单个动作识别方面的准确性和速度,结果显示视觉模块具有最高的准确率和最快的速度。由于手部具有较大的自由度,实现高交互性和实时性是困难的。目前的研究主要集中在人类手臂的跟踪应用和机械臂上的应用,对手部和手指部分的研究较少。

因此,本研究的主要目的有 2 个方面:

(1)针对手指部分,重点选择单根手指的相关角度特征,测试具有自动检测、识别和跟踪功能的系统。

(2)采用 Python 语言作为整体开发工具,以解决跨平台兼容性问题,便于建立基于卷积神经网络的模型,从单色图像中回归手部关节坐标。

综上所述,这是首次综合运用识别、检测、分类和跟踪技术来评估整个系统的性能,并复现手指关节的角度状态。

在当新冠肺炎疫情反复延缓的情况下,该系统有望实现通过手势远程控制医疗器械,并进行无接触式的外科手术操作。此外,通过跟随人体手部动作实现简单手指变化,可以强效辅助病人手部恢复屈曲和伸展功能,以便观察跟随后的手部效果。更进一步地,具备检测、识别、分类和跟踪功能的跟随系统与传统的动态手势图像识别有所不同,不再只输出单一动作,而是可以根据用户的意愿灵活输出所需的手势状态。

1 相关技术解析

1.1 计算机视觉识别手势关键技术

计算机视觉识别手势跟随基本过程如图 1 所示,其核心关键技术为手势分割、手势分析以及手势识别。

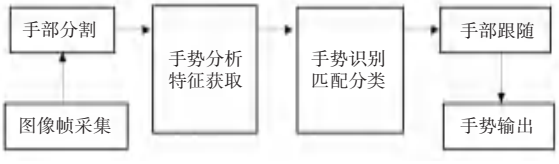


图 1 跟随过程

Fig. 1 Following process

在进行手势识别之前,关键的一步是对所获取的图像进行预处理,以准确地分割手部和背景,其中包括噪声去除、灰度化和边缘检测等处理方法[7],这些预处理步骤能够提高后续手势分析和识别的准确性。其次,手势分析的技术路线和研究方法为手势识别奠定了基础,利用手势的形状和运动轨迹,分析根据何种特征确定手势的姿态,主要的方法有多特征融合法和指关节跟踪法。多特征融合法利用多个特征的组合来描述手势,从而提高识别准确性。指关节跟踪法则专注于跟踪手部指关节的运动信息,以获得手势的姿态。最终,手势识别是一种将模型参数空间里的轨迹或点分类到该空间的某个子集的过程。常用的手势识别方法[8]包括模板匹配法和隐马尔可夫模型法。模板匹配法[9]将手势的动作视为由静态手势图像组成的序列,并将待识别的手势模板序列与已知的手势模板序列进行比较,以识别出手势。而隐马尔可夫模型法[10]则是一种基于统计模型的方法,通过建立双重随机过程来对手势进行建模,其中包括状态转移和观察值输出的随机过程。这些方法和技术为实现准确的手势识别提供了基础,并在不同应用领域发挥着重要作用。

1.2 手部姿态估计算法技术

Google 研发了一个 MediaPipe 的开源项目,可用于时间序列数据处理,具有手掌检测与坐标预测的功能[11],这就为本文研究带来了显著优势,具体机器学习框架如图 2 所示。

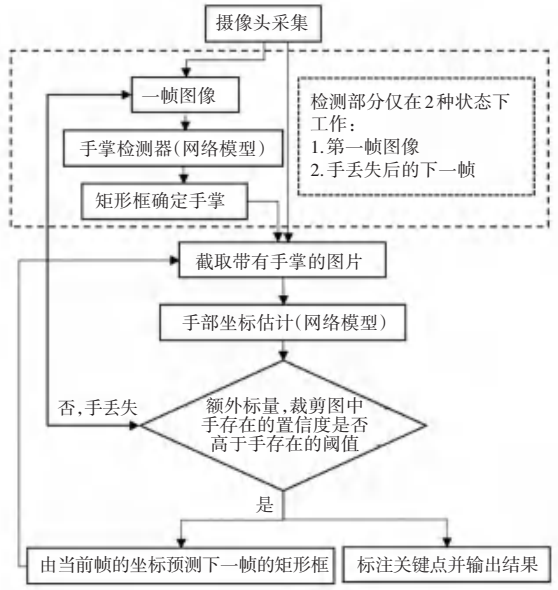


图 2 MediaPipe 机器学习框架

Fig. 2 Machine learning framework of MediaPipe

其中,检测手部使用类似单次检测多目标神经网络(Single Shot Multibox Detector, SSD)[12],能够用较高的准确率和速度检测出图像中手掌的位置和大小。

同时,MediaPipe 利用 3 个操作来优化模型:

NMS、encoder－decoder feature extractor 和 focal loss 来优化模型。其中，NMS 可以抑制算法识别到重复框，得到最高置信度的检测框；encoder－decoder feature extractor 可以对更大的场景上下文进行感知，并识别小对象；focal loss 可以解决正负样本不平衡的问题，有效提高模型的精度。MediaPipe 利用大量标注数据学习手势和姿态的特征和规律，具有更高的准确性和稳定性。

2　系统搭建与设计

2.1　整体框架设计

由于本文重点针对单根手指实现跟随系统，选定便于点对点控制的关键点坐标作为本文图像中的特征点。整个系统分为三大模块：

（1）识别感知，用笔记本本身的摄像头进行信息采集；

（2）跟随决策，则基于理论知识，建立一个可以直接从图像中预测坐标点的简单网络模型，将输入图像的像素直接映射为各点坐标位置；

（3）输出控制，是在 MediaPipe 训练好的模型下获得坐标数据，结合单片机来实现实时驱动舵机转动。总体系统框架如图 3 所示。

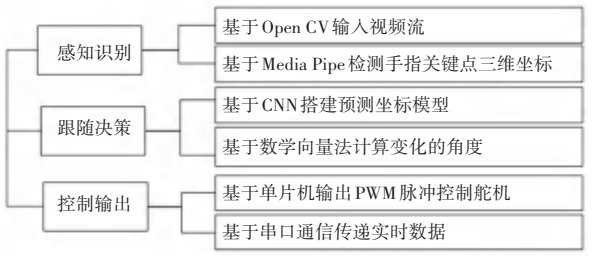
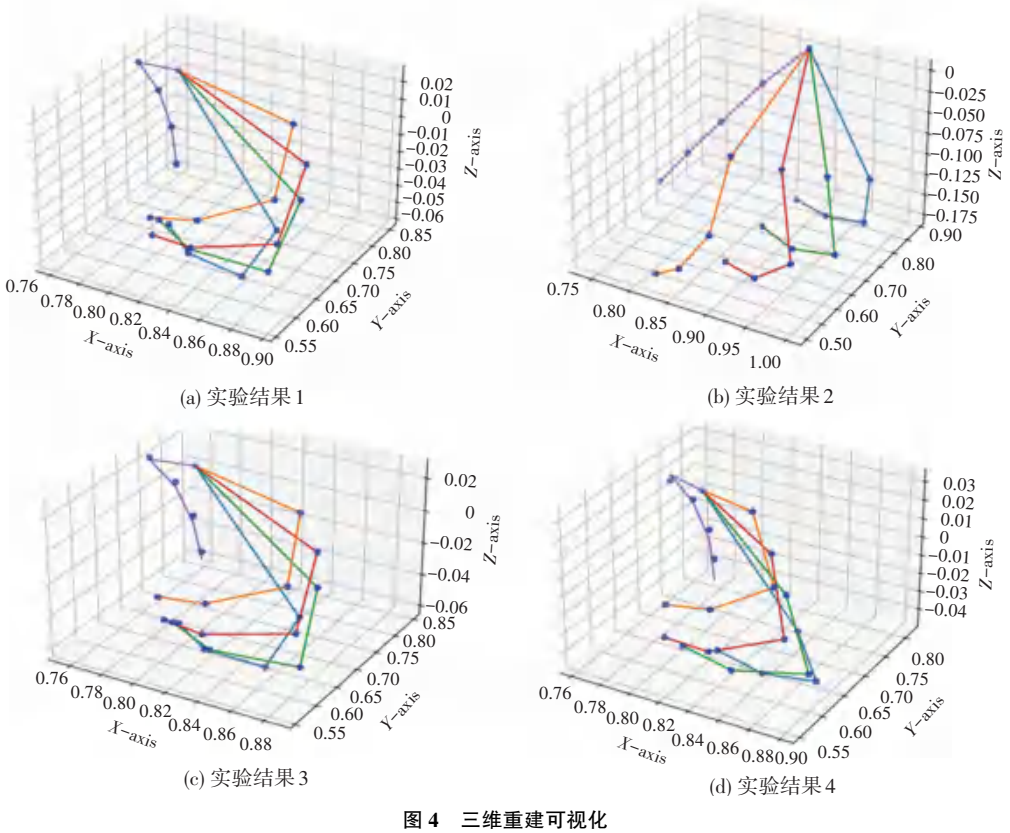


图 3　系统总体框架

Fig. 3　General framework of the system

2.2　感知识别模块搭建与设计

本模块增加公开 Frei Hand 数据集进行后续实验验证，提供了一个大规模、真实采集的手部图像数据集，将数据集划分为训练集与测试集[13]，可用于计算机视觉领域的手势识别、手势控制、手部姿态估计和手势跟随等主题的研究与开发。所有手部图像均通过一个高分辨率的相机进行拍摄，包含 130 240 张图像，对应的相机内部参数矩阵数据和相关的三维坐标值数据标签，包括了遮挡、折叠和扭曲等姿势。此外，为了让计算机获取到实时的手势，直接利用笔记本摄像头对手部信息进行采集，进一步验证系统的准确率和执行速度。其中，采集到的一部分手部信息经三维重建后的效果如图 4 所示。



(a) 实验结果 1　(b) 实验结果 2

(c) 实验结果 3　(d) 实验结果 4

图 4　三维重建可视化

Fig. 4　3D reconstruction visualisation

2.3 跟随决策模块搭建与设计

理论上说,可以在不进行手部检测的情况下直接使用图像或视频数据进行手部姿态预测,以便用于后续的数据输出和控制跟随。因此,本模块基于 CNN 建立一个深度学习回归模型预测图像中手部关键点的位置。

首先定义模型的前向传播函数,将输入的数据传送至第 1 个卷积层,然后经过 ReLU 激活函数,输出特征图。接着经过第 2 个卷积层,再经过 ReLU 激活函数。其次,经过最大池化层进行下采样,特征图的尺寸减半,特征数量加倍。接下来依照相同的模式又经过 4 个卷积层和 2 个平均池化层的处理,进行特征提取。最后,通过 2 个全连接层进行处理,将特征图转换成 1 个一维的向量,并经过 ReLU 激活函数处理后输出。其中,卷积核大小设置为 3×3,步长为 1,填充为 1,最大池化 Max Pooling 层的大小为 2×2,采用 Adam 优化器。

与分类问题不同,关键点坐标预测是一个回归问题,本文参考平均坐标误差的计算方法,把均方误差作为系统评价指标。因此,使用 *MSE* 来计算误差,而不是 Cross Entropy,逐次增加迭代次数和训练轮次以及改变维度,构造多个对比组,在测试集中观察模型预测效果。在每个训练轮次结束后计算误差,训练时间过久会导致过拟合的情况出现,在验证集上未能获得很好的效果,因此,提前停止可以避免模型过度学习训练数据的特点。训练过程的误差评价指标如图 5 所示,应根据实际问题和数据的特点选择合适的停止策略。

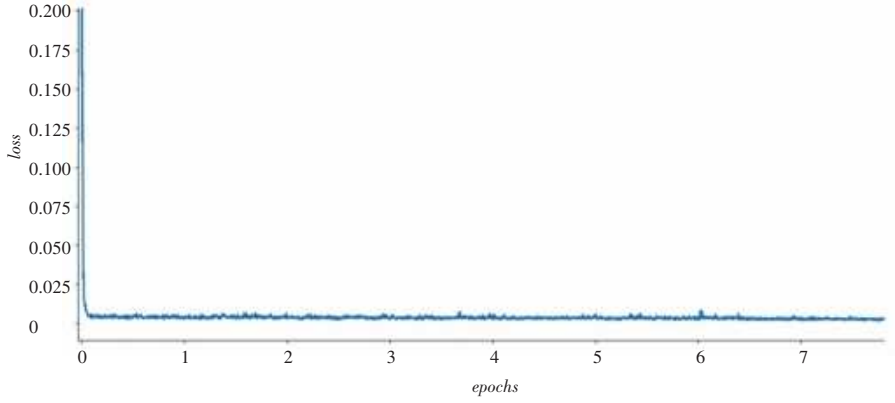


图 5 训练集误差变化

Fig. 5 Training set error variation

2.4 输出控制模块搭建与设计

实物舵机如图 6 所示。图 6 中,红色线接电源,灰色线接地,橙色线接提供脉冲信号的引脚。本文实验采用的舵机是一种位置伺服驱动器,能控制转动 180°以内的角度,适用于那些需要不断变化并可以保持的驱动器中,例如机器人的关节等。

图 6 舵机实物

Fig. 6 The actual servo

本文将所控制的一个舵机视为单根手指上的一个关节点自由度,单根手指处于一个平面上,根据图 7 所展示的几何平面图设计来计算角度,计算公式如下:

$$\begin{cases}\alpha_1 = \arctan\left(\dfrac{y_2 - y_1}{x_2 - x_1}\right)\\ \alpha_2 = \arctan\left(\dfrac{y_3 - y_2}{x_3 - x_2}\right)\\ \theta = 180° - \alpha_1 - (180° - \alpha_2) = \alpha_2 - \alpha_1\end{cases} \tag{1}$$

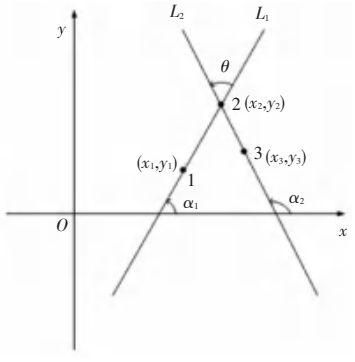


图 7 坐标系

Fig. 7 Coordinate system

根据立体空间上识别到的三维坐标，进一步改进角度计算方法，针对单一手指所构成的平面，在 Python 中计算处于立体空间下由 3 个点构成的 2 条直线向量之间的角度，计算公式为：

$$\cos\theta = \frac{(x_1 \times x_2 + y_1 \times y_2 + z_1 \times z_2)}{\left[\sqrt{(x_1^2 + y_1^2 + z_1^2)} \times \sqrt{(x_2^2 + y_2^2 + z_2^2)}\right]} \tag{2}$$

3　实验与结果

3.1　系统实验方案与环境

本文系统设计的特性是 Win11 电脑端和 ARDUINO 端通信的性质，因此针对系统算法性能测试与控制功能测试的需求。将测试内容分为 Win11 电脑端和 ARDUINO 控制端两部分，实验方案见表 1，并对结果进行展示。

表 1　实验方案

Table 1　Experimental programme

名称	实验内容
Win11 电脑端	MediaPipe 运行时间计算实验 网络模型训练实验
ARDUINO 控制端	客户端信号接收实验 输出结果与显示实验

本文实验所用的硬件设备包括一台笔记本电脑、SG90 舵机和 ARDUINO UNO R3 主板。采用的软件分为 2 部分：Win11 端和 ARDUINO UNO R3 主板 PWM 引脚端。Win11 端深度学习采用的程序语言和软件开发库为 Python、MediaPipe、Pytorch；运行环境需安装 Pycharm2022 软件，以导入相关软件库编译 Python3.8.8 程序文件代码。Arduino PWM 引脚端处采用的程序语言为基于 C/C++编程语言的 Arduino 语言和 Arduino 库控制舵机；运行环境需安装 Arduino IDE，以便运行代码。软件开发库详情见表 2。

表 2　软件开发库

Table 2　Software development library

名称	作用
OpenCV-Python 3.4.1865	实现图像处理和计算机视觉方面的计算机视觉库
Torch 2.0.0	Python 用于建立深度学习模型
MediaPipe 0.9.1.0	检测手部关键点的机器学习算法
Numpy 1.23.5	用于数学计算，可处理大型矩阵

3.2　系统评估指标

评估手势跟随系统的性能涉及多个方面。

(1)平均坐标误差。采用预测的坐标和真实的坐标之间的欧几里得距离表示，是指在欧几里得空间中两点之间的距离，也就是两点直线距离(即勾股定理)[14]。

在三维空间中，两点(x_1,y_1,z_1)和(x_2,y_2,z_2)之间的欧几里得距离为：

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{3}$$

(2)实时性。系统能否及时响应用户手势通常可以通过程序执行所需的时间来评估，即处理一帧图像的时间与处理前的时间之差 T：

$$T = T_2 - T_1 \tag{4}$$

其中，T_2 表示处理后的时间，T_1 表示处理前的时间。

每秒可以处理的帧数 FPS 可由下式计算求出：

$$FPS = \frac{1}{T} \tag{5}$$

(3)兼容性。通过实验判断手势跟随系统是否能够与其他应用程序和设备进行跨平台交互，测试交互性并确保系统与其他应用程序和设备的兼容性，对于用户体验非常重要。

(4)均方误差(Mean Squared Error, MSE)。在回归问题中使用，计算预测值和真实值的差值的平方：

$$MSE = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2 \tag{6}$$

3.3　对比实验结果分析

在实验图片下进行训练测试，展示多个对比实验组的可视化效果，得出以下结论：

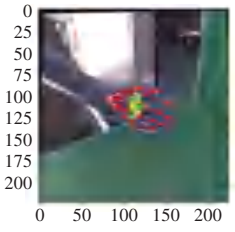
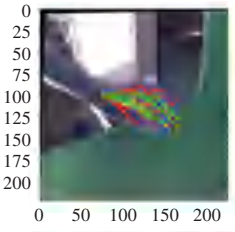
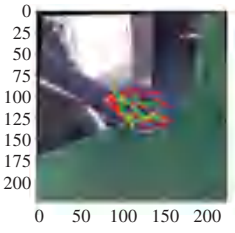
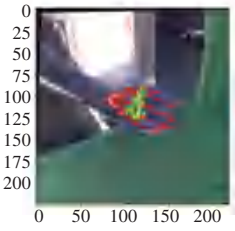
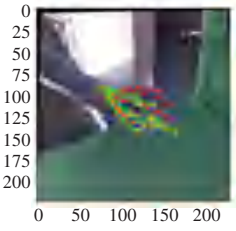
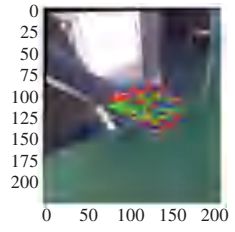
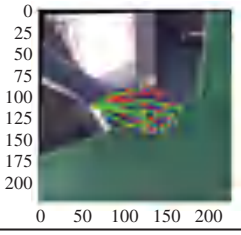
(1)CNN 随着训练次数的叠加，测试时的坐标预测结果值与实际值的误差变小，损失函数的值也随之越来越小。

(2)由于笔记本 CPU 的性能较低，选择合适的数据集中用于训练的数量，在经过一定的训练轮次后，能够达到均方误差的最低值。

(3)学习率直接影响模型的收敛快慢，过大会错失信息、过小则无法收敛。收敛速度的快慢不重要，找到一个合适的训练结果才是重要的。对比二维和三维坐标的预测后发现，可视化结果相差不大。由此可见，在已标注了三维坐标的图像中，可将其作为样本用于深度学习，便于后续实际情况中直接输出三维坐标。研究得到的训练效果见表 3。

表 3 训练效果

Table 3 Training effect

维度	训练次数(*epochs*)	效果	维度	训练次数(*epochs*)	效果
二维	1		二维	30	
二维	5		三维	1	
二维	10		三维	10	
二维	20				

4 系统跟随可行性验证

本文将 MediaPipe 库中输出的关键点坐标数据转化为角度,进行点对点控制,验证通过串口给开发板传输角度数据具有非常高的实时性。减少不必要的程序后,将 *FPS* 从 10 提升到了 23。利用 Arduino 软件来对单片机发送控制命令,在计算机与单片机之间建立串口通信连接,通过应用库函数实现角度数据的连续传送与获取,实现单片机控制舵机转动相应角度。现有设备仅为 1 台舵机,对应图 8 中的 6 号点。

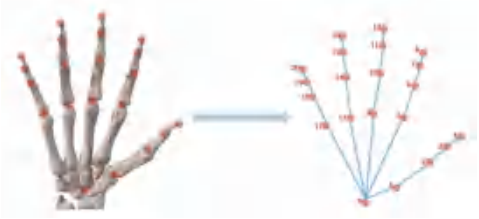


图 8 手部关键点划分

Fig. 8 Division of key points of the hand

采用量角器测量手指该点屈曲伸展的角度,记录了程序预测的角度,再测量舵机转动的角度,最终得到误差与响应速度的实验结果(见表 4)。

表 4 角度实验结果

Table 4 Results of angle experiments

真实角度/(°)	预测角度/(°)	舵机转动角度/(°)	误差/(°)	响应速度/fps
180	172	172	8	12
125	132	132	7	15
90	110	110	10	11

Arduino 开发板只需要指定串口的波特率,就可以接收来自串口的数据,可移植性强,不依赖于笔记本的性能。在视觉上将舵机看作人每段手指之间连接的转动副,即一个自由度,运用数学工具,例如线性变换、向量法等计算手指各转动副处于当前帧的空间转动角度信息,循环更新数据列表中的数据,通过串口通信及时传递到 Arduino 开发板上,控制相应序号的舵机进行转动,模拟人体单根手指的关节旋转情况,实现结果如图 9 所示。图 9 展示手部进行不同运动状态时转动副处角度的实时变化。

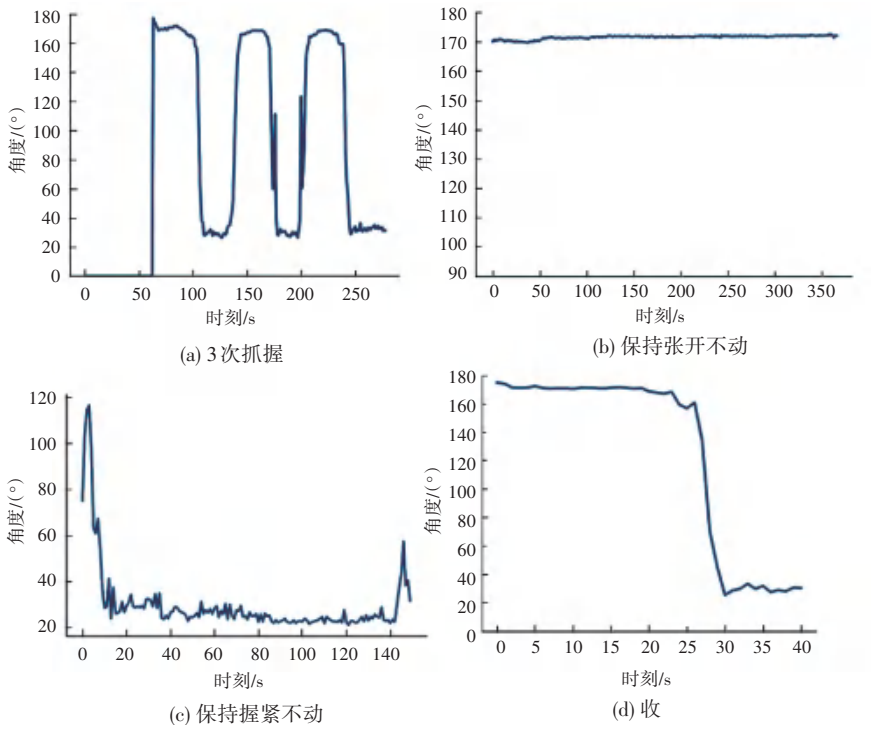


图 9　手部变化

Fig. 9　Hand changes

5　结束语

本文结合现有的先进技术以及对卷积神经网络研究，搭建包含手势的检测、识别和跟踪功能的手势自动跟随系统。首先，借助于 MediaPipe 手势三维坐标的输出，在三维立体空间对角度进行了数学方法计算。其次，控制一个自由度（关键点）的输出，最后实现对手势的自动跟随。实验证明在手势跟随的自动方面达到了一定的实时性能。本研究以实验法为基础，创新性地探索从输入手势到舵机响应运动的整个过程，将视觉识别技术与手势自动跟随相结合应用于实际应用领域，扩展人机姿势协同运作，有助于更好地理解人体运动的规律和特征，具有非常重要的理论和实用价值。

参考文献

[1] 杨志文. 基于表面肌电信号的动态手势识别研究[D]. 武汉：武汉科技大学，2022.

[2] KONG Fanchen, DENG Jingcheng, FAN Zichuan. Gesture recognition system based on ultrasonic FMCW and ConvLSTM model[J]. Measurement, 2022, 190: 110743.

[3] 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.

[4] 詹金峰. 基于机器视觉的手势检测与跟踪技术研究[D]. 广州：华南理工大学，2021.

[5] 张继凯，李琦，王月明，等. 基于单目 RGB 图像的三维手势跟踪算法综述[J]. 计算机科学，2022，49(4)：174-187.

[6] XIAO Feiyun, ZHANG Zhen, LIU Changhai, et al. Human motion intention recognition method with visual, audio, and surface electromyography modalities for a mechanical hand in different environments[J]. Biomedical Signal Processing and Control, 2023, 79: 104089.

[7] 李宁. 手眼伺服作业机器人平台系统的研究[D]. 秦皇岛：河北科技师范学院，2017.

[8] 王慧. 智能机器人静动态手势设计及识别[D]. 呼和浩特：内蒙古大学，2020.

[9] 武霞，张崎，许艳旭. 手势识别研究发展现状综述[J]. 电子科技，2013，26(6)：171-174.

[10] 蔡小龙. 增强现实中的手势交互系统研究与设计[D]. 重庆：重庆大学，2018.

[11] ZHANG Fan, VALENTIN B, ANDREY V, et al. MediaPipe hands: On-device real-time hand tracking[J]. arXiv preprint arXiv，2006. 10214, 2020.

[12] VALENTIN B, KARTYNNIK Y, VAKUNOV A, et al. BlazeFace: Sub-millisecond neural face detection on mobile GPUs[J]. arXiv preprint arXiv，1907. 05047v1, 2019.

[13] CHRISTIAN Z, CEYLAN D, YANG Jimei, et al. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2019: 813-822.

[14] 王丽萍，汪成，邱飞岳，等. 深度图像中的 3D 手势姿态估计方法综述[J]. 小型微型计算机系统，2021，42(6)：1227-1235.