

# 基于改进 YOLOv5+Kalman 的动态手势识别跟踪算法研究

张睿敏, 杜叔强, 周秀媛

(兰州工业学院计算机与人工智能学院, 甘肃 兰州 730050)

✉ 371247813@qq.com; 42787413@qq.com; 46652899@qq.com



**摘要:**针对在边缘端算力较弱的设备上,动态手势检测模型较大、检测时间较长的问题,提出了以 YOLOv5 网络为基础的动态手势识别跟踪算法。首先在 YOLOv5 网络模型上添加自适应注意力模块输出大小相同的张量,添加特征增强模块弥补高层特征层的信息损失,提高特征金字塔的表示能力。其次使用 Kalman 滤波器预测机,形成有预测机制的动态手势识别跟踪算法。预测跟踪时,把短时的手部运动粗略地看成是一种恒定的运动,即在设定的非常短的时间内目标的状态是不变的,也就是相邻两帧图像状态差异不大,所以手势的运动模型采用的是等速度运动模型。利用大量正负样本训练模型,证明算法能实时地对 13 种动态手势进行识别和跟踪,识别准确率达到 99.6%。

**关键词:**改进 YOLOv5;动态手势;Kalman;预测跟踪

**中图分类号:**TP391 **文献标识码:**A

## Research on Dynamic Gesture Recognition and Tracking Algorithm Based on Improved YOLOv5 + Kalman

ZHANG Ruimin, DU Shuqiang, ZHOU Xiuyuan

(College of Computer and Artificial Intelligence, Lanzhou Institute of Technology, Lanzhou 730050, China)

✉ 371247813@qq.com; 42787413@qq.com; 46652899@qq.com

**Abstract:** Aiming at the problems of large dynamic gesture detection models and long detection time on devices with weak computing power at the edge, this paper proposes a dynamic gesture recognition and tracking algorithm based on YOLOv5 network. Firstly, an adaptive attention module with the same output size tensor is added to the YOLOv5 network model and a feature enhancement module is added to compensate for the information loss in the high-level feature layer, so as to improve the representation ability of the feature pyramid. Then, the Kalman filter predictor is used to form a dynamic gesture recognition tracking algorithm with a prediction mechanism. When predicting the tracking, short-term hand movements are roughly regarded as a constant motion, meaning that the target's state remains unchanged for a set very short period of time. That is to say, there is not much difference in the states of adjacent frames of images. Therefore, the motion model of the gesture adopts an equal velocity motion model. The experiment utilizes a large number of positive and negative samples to train the model, proving that the proposed algorithm can recognize and track 13 dynamic gestures in real-time, with a recognition accuracy of 99.6%.

**Keywords:** improved YOLOv5; dynamic gesture; Kalman; predictive tracking

## 0 引言 (Introduction)

随着计算机视觉技术的发展,基于摄像头的手势识别研究受到了学习和业界的重视,手势识别在智能人机交互、智能驾驶系统等方面都有着广阔的应用前景<sup>[1]</sup>。高效的动态手势识

别算法能让计算机准确地识别复杂的手部动作,张建荣<sup>[2]</sup>提出的动态时间规整(improve Dynamic Time Warping, iDTW)手势识别算法,采用点和线结合的方法约束搜索路径,提高了手势识别的精确度。KHARI 等<sup>[3]</sup>使用微调后的 VGG19(Visual

Geometry Group)模型对 RGB(Red Green Blue)和 RGB-D (RGB + Depth) 图像进行手势识别, 识别准确率达到 94.8%, 但只能对静态手势进行识别。王龙等<sup>[4]</sup>通过建立手势肤色模型, 提取手势特征, 能快速识别相关手势, 但是容易受到光照等因素的影响。丁驰等<sup>[5]</sup>提出将神经网络和 SSD(Single Shot Multibox Detector)结构结合起来提取手势特征, 提高了在光线等因素影响下手势的识别效率, 但模型较大、检测时间长。

从以上学者的研究可以看出, 动态手势识别算法已经取得了不错的识别效果, 但研究者很难设计出一种对所有手势都适用的识别方法。基于改进 YOLOv5+Kalman 的动态手势识别跟踪算法主要在 YOLOv5 网络的基础上增加了两个模块: 自适应注意力模块(Adaptive Attention Module, AAM)和特征增强模块(Feature Enhancement Module, FEM)。增加这两个模块后, 一方面减少了生成特征图时的信息丢失问题, 另一方面增强了特征金字塔的表示能力, 以便对不同手势进行实时准确的识别, 然后应用 Kalman 滤波器进行预测跟踪。预测跟踪时, 将短时内的手部运动看作恒定的运动, 目标状态在设定的非常短的时间内是不变的, 在跟踪时, 因为相邻两帧图像差异不大, 所以手势的运动模型采用的是等速度运动模型。

## 1 改进 YOLOv5 网络(Improved YOLOv5 network)

### 1.1 YOLOv5 网络模型

YOLOv5 模型是基于 YOLOv3 模型的基础上改进而来的, 目前有 YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x 四个模型。YOLOv5 的组成为三个部分: Backbone(骨干网络)、Neck(颈部网络)、Output(头部网络), 具体如图 1 所示。

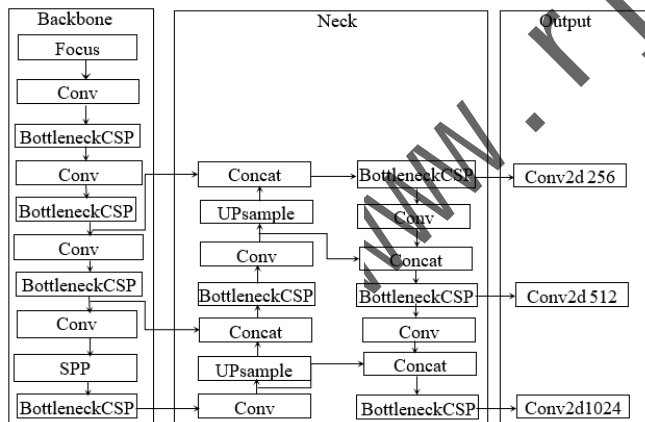


图 1 YOLOv5 网络模型

Fig. 1 YOLOv5 network model

Focus、Conv、BottleneckCSP 和 SPP 等模块组成模型的 Backbone 网络, Backbone 主要是在不同图像细粒度上聚合并提取图像特征。YOLOv5 通过在骨干网络 Backbone 和 Neck 网络部分使用 BottleneckCSP 模块, 在保证语义提取准确的同时, 提高网络速度<sup>[6]</sup>。Neck 网络组合了一系列图像特征, 并将图像特征传递到 Output 网络。Output 网络对图像进行检测并生成边界框和预测类别。设置大小为  $80 \times 80 \times 512$  的图像, Neck 网络进行上采样操作, 上采样操作包含两个瓶颈层

BottleneckCSP, 两个标准卷积层 Conv(大小步长为 1), 两个 Upsample 和 Concat 层; 然后进行下采样操作, 下采样操作包括两组 BottleneckCSP、Conv、Concat 层和单独的 BottleneckCSP, 得到大小分别是  $80 \times 80 \times 256$ 、 $40 \times 40 \times 512$ 、 $20 \times 20 \times 1024$  的特征图。

YOLOv5 应用大小不同的特征图在头部进行检测, 即应用大图检测小目标, 应用小图检测大目标。对于颈部操作得到的结果, 再经 Conv2d 卷积操作, 便得到大小为  $80 \times 80 \times 256$ 、 $40 \times 40 \times 512$ 、 $20 \times 20 \times 1024$  的特征图。

### 1.2 AF-YOLOv5 网络

在 YOLOv5 原有特征金字塔的基础上, 添加 AAM 和 FEM 两个模块, 既减少了生成特征图时的信息丢失问题, 也增强了特征金字塔的表示能力, 如图 2 所示。这样在保证实时检测的前提下提高了 YOLOv5 网络对多尺度目标的检测性能, 各组分别采取具有差异卷积核的大小卷积, 从而获取不同尺度的感受野<sup>[7]</sup>, 得到不同尺度的目标信息。此外, 增加了自动学习数据增强模块, 在增强数据集的同时提高了模型的鲁棒性, 使其更适用于实际场景。

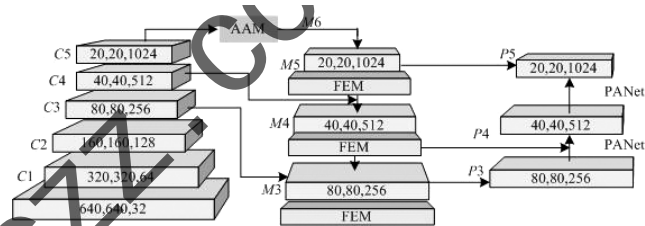


图 2 AF-YOLOv5 网络

Fig. 2 AF-YOLOv5 network

输入图像经过多个卷积操作生成  $\{C1, C2, C3, C4, C5\}$  特征映射。C5 通过 AAM 生成特征映射  $\{M6\}$ , 映射  $\{M6\}$  与  $\{M5\}$  求和, 然后传递给下一层与其他特征融合, 通过有限元分析, 路径聚合网络(Path Aggregation Network, PANet)缩短了底层与顶层特征间的信息路径。首先, AAM 模块通过自适应池化层得到不同尺度的语义特征; 然后, 每个上、下文特征进行  $1 \times 1$  卷积, 以获得通道维数, 利用双线性插值将其上采样, 得到 S 尺度以便进行后续融合, 空间注意力机制通过 Concat 层将三个上、下文特征的通道进行合并; 特征图依次经过  $1 \times 1$  卷积层、整流线性单元(Rectified Linear Unit, ReLU)激活层、 $3 \times 3$  卷积层和 Sigmoid 激活层, 为每个特征图生成相应的空间权重。

FEM 模块由多分支卷积层和分支池化层两个部分组成, 多分支卷积层能为输入特征图提供不同大小的感受野, 它通过扩张卷积、批归一化(Batch Normalization, BN)层和 ReLU 激活层。把三个平行分支中的扩张卷积核设定为相同大小, 都为  $3 \times 3$ , 扩张速率设定为不同, 分别为 1、3、5。分支池化层用于融合来自不同并行分支的信息, 避免引入额外参数<sup>[8-9]</sup>。在训练过程中, 不同分支的平衡通过平均操作实现, 使单个分支在测试过程中实现推理。

### 1.3 数据增强预处理

图像增强通过一定手段对原图像附加一些信息或变换数据, 有选择地突出图像中感兴趣的特征或者抑制(掩盖)图像中某些不需要的特征, 使图像与视觉响应特性相匹配<sup>[10-12]</sup>。图像

数据增强就是对被训练的图像做随机改变(如旋转、缩放、变形等),目的是生成相似而又不同的训练样本,进而扩大训练数据集大小。模型通过多种类型数据集的训练,能够降低模型对有些特定属性的依赖,提高模型的泛化能力。同时,图像本身的变化能够增强模型对未见数据的泛化,从而避免了模型的过度拟合。

本文增强图像时使用的是 ImageDataGenerator() 类, ImageDataGenerator() 类是图片生成器,它包含在 keras.preprocessing.image 包中;每次使用它对 batch\_size 个样本图像进行增强,部分关键代码如下:

```
(x_train, y_train), (x_test, y_test) = tf.keras.datasets.mnist.load_data() //获取数据集
x_train = X_train.reshape(X_train.shape[0], 28, 28, 1) //将数据转换为 4 维的形式
x_test = X_test.reshape(X_test.shape[0], 28, 28, 1)
datagen = ImageDataGenerator(horizontal_flip=True) //设置图像增强方式:水平翻转
for X_batch, y_batch in datagen.flow(x_train, y_train, batch_size=9):
    //查看增强后的结果
    plt.figure(figsize=(8, 8)) //设定每个图像显示的大小
    for i in range(0, 9): //产生一个 3*3 网格的图像
        plt.subplot(330+1+i)
        plt.title(y_batch[i])
        plt.axis('off')
        plt.imshow(X_batch[i].reshape(28, 28), cmap='gray')
    plt.show()
    break
```

## 2 Kalman 预测跟踪模型(Kalman predictive tracking model)

在跟踪算法上加一个预测机,本文使用 Kalman 滤波器预测机<sup>[13-14]</sup>,形成有预测机制的跟踪算法。把短时的手部运动粗略地看成一种恒定的运动,在设定的非常短的时间内目标的状态是不变的,设计 Kalman 预测跟踪机对手势进行预测跟踪。通过遍历图像特征的所有通道,减去平均像素值,对于去均值后的特征图,再遍历所有特征<sup>[15-17]</sup>,得到  $S, V, D(x_i = VSD)$ ,取  $V$  矩阵前  $K$  列,即对应特征的压缩维度(预设值)。在预测跟踪中,手势运动模型采用的是等速度运动模型,状态向量定义为公式(1):

$$X(t) = [colcenter(t), rowcenter(t), v_x(t), v_y(t)]^T \quad (1)$$

观测向量定义为公式(2):

$$Z(t) = [colcenter(t), rowcenter(t)]^T \quad (2)$$

预测向量定义为公式(3):

$$X'(t) = [colcenter'(t), rowcenter'(t), v'_x(t), v'_y(t)]^T \quad (3)$$

其中,  $colcenter'(t), rowcenter'(t)$  表示 Kalman 滤波器预测的搜索窗口的中心位置,  $v'_x(t), v'_y(t)$  表示搜索速度。

将 Kalman 滤波的状态方程定义为公式(4):

$$X_{t+1} = A_t X_t + B_t u_t + w_t \quad (4)$$

观测方程定义为公式(5):

$$Y_t = Hx_t + v_t \quad (5)$$

以上方程中,参数  $w_t$  是  $t$  时刻的动态噪声,  $v_t$  是  $t$  时刻的观测噪声,其余参数都表示相关矩阵,  $x_t$  表示状态,  $u_t$  表示控制,  $Y_t$  表示  $t$  时刻的观测,  $A_t$  表示状态转移,  $B_t$  表示状态控制,  $H$  表示观测。

将时间更新方程定义为公式(6):

$$\begin{cases} \hat{x}_t^- = A\hat{x}_{t-1}^- + B_t u_{t-1} \\ P_t^- = A P_{t-1}^- A^T + Q \end{cases} \quad (6)$$

将状态更新方程定义为公式(7):

$$\begin{cases} K_t = P_t^- H^T (H P_t^- H^T + R)^{-1} \\ \hat{x}_t = \hat{x}_t^- + K_t (Y_t - H \hat{x}_t^-) \\ P_t = (I - K_t H) P_t^- \end{cases} \quad (7)$$

以上方程中,  $P_t$  是协方差矩阵,  $\hat{x}_k$  是新的最优估计,  $K_t$  是卡尔曼增益,它表明当前时刻新的最优估计是由上一时刻的最优估计加上外部控制量预测得到的。这样 Kalman 滤波器预测模型通过时间更新和状态更新保证跟踪的实时性和准确性。

## 3 实验过程与结果(Experimental process and results)

### 3.1 实验环境部署

实验硬件配置要求:CPU 为 AMD Ryzen 5 3600 @4.2 GHz, GPU 为 NVIDIA GeForce RTX 3060,内存为 32 GB,显卡为 ASUS TUF-GeForce RTX3080TI-O12G-GAMING 12 GB 显存。软件开发工具:官网下载安装 Python3;安装 Python 编辑器 PyCharm;安装 Anaconda;打开 cmd 窗口输入命令 pip install pandas,安装 pandas 库;打开 cmd 窗口输入命令 pip install opencv,安装 OpenCV 库;官网下载安装 CUDA;官网下载安装 cuDNN;安装 PyTorch。

### 3.2 训练集

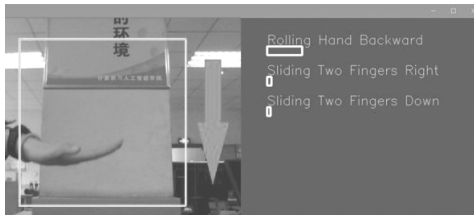
利用大量样本训练模型,实验数据集采用动态手势图像,其来源主要包括向前、向后、向左、向右、点赞等共 13 种动态手势数据,包括 600 张正样本(手势)和 600 张负样本(不出现手势的背景),负样本在一定程度上增加了模型的鲁棒性和泛化功能。正负样本的比例为 1:1,用来进行模型训练,借助开源计算机视觉库(Open Source Computer Vision Library, OpenCV)进行对应的手势模型训练,然后进行手势识别跟踪训练。

### 3.3 实验结果

动态手势识别跟踪系统采用改进的 YOLOv5 卷积神经网络模型+Kalman 预测跟踪模型进行动态手势的识别和跟踪,通过对算法的改进可以很好地提升手势识别精度,识别准确率达到了 99.6%,而且不会占用太多资源。系统对部分手势识别跟踪的具体运行结果如图 3 所示。

在项目运行结果窗口中,左侧窗口通过检测框和箭头的组

合表示对手势的识别和手势的运动方向,右侧窗口通过文字和滚动状态框显示识别到的动态手势结果。其中,图 3(a)的识别结果是“Rolling Hand Backward”,图 3(c)的识别结果是“Swiping Left”。



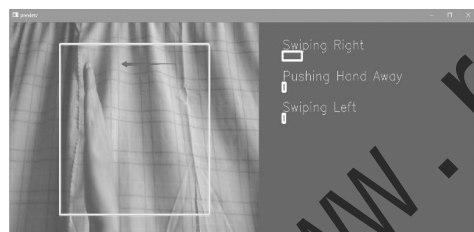
(a)向后挥动



(b)向前挥动



(c)向左挥动



(d)向右挥动

图 3 运行结果

Fig. 3 Operation results

## 4 结论(Conclusion)

本文针对动态手势的识别跟踪问题,提出基于改进的YOLOv5 卷积神经网络模型+Kalman 预测跟踪模型进行动态手势的识别和跟踪算法。算法是在 YOLOv5 原有的特征金字塔网络基础上,增加了自适应注意力模块 AAM 模块和特征增强 FEM 模块,起到了减少特征图生成过程中信息丢失并增强表示能力的作用,设计 Kalman 跟踪器,通过时间更新和状态更新,能适时、准确地对动态手势进行跟踪。采用大量样本进行实验,实验证明算法能够对基本动态手势进行快速且准确的识别,并且 Kalman 滤波器预测模型保证了对动态手势跟踪的实时性和准确性。为了进一步提高算法对更多动态手势的识别和跟踪,后续将对算法进行持续的优化研究。

## 参考文献(References)

- [1] 徐访,黄俊,陈权. 基于 3D 卷积神经网络的动态手势识别模型[J]. 计算机工程,2021,47(11):283-291.
- [2] 张建荣. 基于 Kinect 手势识别的虚拟环境体感交互技术研究[D]. 重庆:重庆邮电大学,2016.
- [3] KHARI M, GARG A K, CRESPO R G, et al. Gesture recognition of RGB and RGB-D static images using convolutional neural networks[J]. International Journal of Interactive Multimedia and Artificial Intelligence, 2019, 5(7): 22-27.
- [4] 王龙,刘辉,王彬,等. 结合肤色模型和卷积神经网络的手势识别方法[J]. 计算机工程与应用, 2017, 53(6): 209-214.
- [5] 丁驰,林军,游俊,等. 基于深度学习的手势识别方法[J]. 控制与信息技术, 2018(6): 96-99.
- [6] 闫颖月,王伟,田泽. 复杂环境下基于改进 YOLOv5 的手势识别方法[J]. 计算机工程与应用, 2023, 59(4): 224-234.
- [7] 彭玉青,赵晓松,陶慧芳,等. 复杂背景下基于深度学习的手势识别[J]. 机器人, 2019, 41(4): 534-542.
- [8] GUO Y L, WANG H Y, HU Q Y, et al. Deep learning for 3D point clouds a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(12): 4338-4364.
- [9] 刘玉鹏. 动态手势识别和跟踪算法研究[D]. 济南:山东大学, 2019.
- [10] 周云,胡锦楠,赵瑜,等. 基于卡尔曼滤波改进压缩感知算法的车辆目标跟踪[J]. 湖南大学学报(自然科学版), 2023, 50(1): 11-21.
- [11] QIAN W, WANG G Z, LI G P. Improved robust algorithm for real-time traffic light detection with YOLOv5[J]. Computer Science and Exploration, 2022, 16(1): 231-241.
- [12] 阎磊,曲全福,白涛,等. 基于扩展卡尔曼滤波的环路跟踪算法仿真分析[J]. 火力与指挥控制, 2022, 47(2): 14-17.
- [13] 韩锐,杨穷千. 融合运动状态信息的高速相关滤波跟踪算法[J]. 湖南大学学报(自然科学版), 2020, 47(4): 82-91.
- [14] 潘迪夫,李耀通,韩锐. 一种基于多相关滤波器组合的目标跟踪方法[J]. 湖南大学学报(自然科学版), 2019, 46(2): 112-122.
- [15] 李晓云,何秋生,张卫峰,等. 自适应多尺度颜色特征目标跟踪算法研究[J]. 液晶与显示, 2019, 34(3): 291-301.
- [16] 汤子麟,刘翔,张星. 光照不均匀图像的自适应增强算法[J]. 计算机工程与应用, 2021, 57(21): 216-223.
- [17] 王敏敏. 基于压缩感知的动目标检测技术及其应用研究[D]. 上海:中国科学院大学(中国科学院上海技术物理研究所), 2018.

## 作者简介:

张睿敏(1978-),女,硕士,副教授. 研究领域:图像识别与数据分析,Web 信息组合.

杜叔强(1978-),男,硕士,副教授. 研究领域:计算机软件与理论.

周秀媛(1976-),女,硕士,副教授. 研究领域:计算机图形图像.