DOI: 10.13382/j.jemi.B2003749

基于 Mask R-CNN 与 SG 滤波的手势识别关键点特征提取方法*

王婧瑶[1,2] 王红军[2,3]

(1. 北京工业大学 人工智能与自动化学院 北京 100124; 2. 北京信息科技大学 高端装备智能感知与控制北京市国际科技合作基地 北京 100192; 3. 北京信息科技大学 机电工程学院 北京 100192)

摘 要:手势识别是人机交互的重要手段。为了精确识别手势并摒除光照等环境干扰,同时减除由于手部高维运动造成的关键点剧烈抖动的问题,提出一种基于基于蒙版区域的卷积神经网络(Mask Region-based convolutional neural network,Mask R-CNN)与多项式平滑算法(Savitzky-Golay,SG)的手势关键点提取方法。该方法首先对输入的红绿蓝(RGB)三通道图像进行特征提取与区域分割,获得手部的实例分割与掩码。然后利用 ROIAling 及功能性网络进行目标匹配,标记出 22 个关键点(21 个骨骼点+1 个背景点)。将标记后结果送入 SG 滤波器进行数据平滑,并进行骨骼点的重新标定。从而得到稳定的手势提取特征。对模型进行对比实验,结果表明,该方法能够最大程度摒除环境干扰,并精准提取关键点。与传统基于轮廓分割的手势关键点提取相比,模型的鲁棒性大大提高,识别精度达到 93.48%。

关键词: 计算机视觉;手势识别;关键点提取;Mask R-CNN;Savitzky-Golay 滤波

中图分类号: TP301;TP391 **文献标识码**: A **国家标准学科分类代码**: 520.20;520.50

Gesture key point extraction method based on Mask R-CNN and SG filter

Wang Jingyao[1] Wang Hongjun[2,3]

(1. College of Artificial Intelligence and Automation, Beijing University of Technology, Beijing 100124, China;
2. Intelligent Sensing and Control of High-end Equipment Beijing International Science and Technology Cooperation Base, Beijing Information Science and Technology University, Beijing 100192, China;
3. School of Mechanical and Electrical Engineering, Beijing Information Technology University, Beijing 100192, China)

Abstract: Gesture recognition is an important means of human-computer interaction. In order to more accurately recognize gestures and eliminate the interference of environmental conditions such as lighting, and reduce the key point jitter recognition error caused by the high-dimensional space transformation of the hand at the meanwhile, a gesture key point method of extraction based on the Mask R-CNN model and Savitzky-Golay filter is proposed. This method uses the Mask R-CNN model to process RGB three-channel images, and performs object recognition and segmentation on each image, and obtains 21 bone points and background positions of the hand and performs model training. Then uses neural network features to match the video stream and mark 22 key points. Furthermore, the point data is smoothed by using Savitzky-Golay filter and then redraw the data to obtain stable gesture extraction and reconstruction results. This method is used in bone point extraction experiments. Experimental results show that the method can eliminate environmental interference to the greatest extent and accurately extract key points. Compared with traditional gesture key point extraction based on contour segmentation, the accuracy reaches 93.48%. At the same time, the robustness of the model is greatly improved.

Keywords: computer vision; gesture recognition; key point extraction; Mask R-CNN; Savitzky-Golay filter

收稿日期:2020-12-07 Received Date: 2020-12-07

*基金项目:国家科技部项目(G20190201031)、北京信息科技大学科研内涵发展重点研究培育项目(2020KYNH226)资助

0 引 言

手势作为人类的基本特征,在人机交互、机械控制、虚拟现实等领域具有重要意义。利用视觉技术,计算机已可以实现手势取代传统输入对机器进行控制,虚拟交互,手语认知等复杂任务。而完成这些任务的基础则是精准提取手部关键点并进行手势识别。传统的使用数据手套、借助加速度传感器、使用特殊标记等方法都无法摆脱繁复外设的束缚,基于视觉的依靠手势区域分割与轮廓提取的方法则在精度与鲁棒性上还存在一些不足。

近年来深度学习与神经网络技术快速发展,将其与传统视觉技术相结合成为了一种新的研究方向。Simonyan 等[1]提出采用双 stream,静态单张图片与多张图分类,初步实现多帧图像的同步手势处理,但容易出现过拟合的问题。Zimmermann 等[2]提出利用正则化对手势进行坐标标定,精度上进行了优化,但实时性较差。Molchanov 等[3]提出的一种端到端的多模态[4]手势识别模型,在 color+depth+optical flow[5]三种数据的测试条件下,平均精度较其他模型大大提高。但对于大数据处理,部分识别结果抖动剧烈,且有较大时间损耗。Hu[6]等提出了一种新型的注意力与序列网络(ASNet)用于准确判定手部关节序列机制,一定程度解决了识别抖动剧烈的问题,但识别速度依旧无法达到理想预期。手势识别综合效果较优的为 Yang[7]等提出的一种用于联合手势识别和 3D 手势估计的新型协作学习网络。基于网络的联合感知功能将手势识别与 3D 手势估计结合起来,精度远超于 20BN-jester 基准测试的最新水平。但该算法会产生不必要的资源浪费。

为了解决上面的问题并进行效果优化,本文采用将 Mask R-CNN[8]神经网络模型和 SG 滤波(Savitzky-Golay filter)相结合的方式,来实现手部骨骼点的识别标注与平滑。Mask R-CNN 是针对单张图像进行物体分割与识别的,通过在 Faster-RCNN[9]的基础上添加一个分支网络,在实现目标检测的同时,把目标像素分割出来。结合图像金字塔网络,对尺度不同物体的识别效果进行优化,并引入全卷积网络来实现精确的实例分割。

为了更精确地识别到特定骨骼点,本文利用 Mask R-CNN 进行位置估计,取代了传统的利用方向梯度直方图+支持向量机(HOG+SVM)[10]、CNN[11]或 SIFT 局部特征描述子[12]的方法,得到更精确的实例分割与标定结果。并利用 SG 滤波器进行了数据平滑。减弱了视频数据流中由于高维运动造成的骨骼点抖动,令手势骨骼点标定算法得以进一步优化。

1 手势骨骼点定位识别模型

1.1 模型原理

本文搭建的系统模型输入为单目实时捕捉图像,进行实时性的位置标注后,将标有 21 个手势关键点的结果进行输出。

方法采用 Mask R-CNN 作前向计算[13],提取图像中的手部信息,获得对手部不同部位进行分割后的特征图。并利用中间层特征,对预设的关键点信息以及处理图像进行匹配,取曼哈顿距离最小的对应点作为识别到的关键点,初步获得标定结果。

Mask R-CNN 是 He 等[8]在 Faster R-CNN 的基础上研发的深度神经网络模型,该模型在识别和分割单张图片中的物体的任务中有着很出色的效果。

实例分割(instance segmentation)算法中,Mask R-CNN 属于第 1 梯队。它有着高处理速度、高精度、简单直观等优势,是一个非常灵活的框架。可以增加不同的分支完成不同的任务,例如目标分类、目标检测、语义分割、实例分割、人体姿势识别[13]等。

本方法所构建的模型,将 Mask R-CNN 结构划分为特征提取与特征组合两部分。并在此基础上,引入区域提交网络,ROIAlign 与功能性网络(包括分类、二段修正、分割)3 层,构建了针对小面积(手部)区域的精确分割与识别模型。

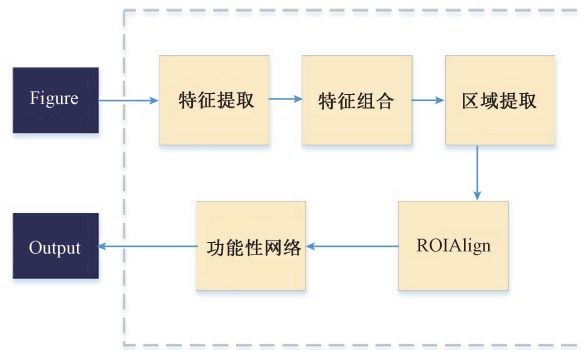


图 1 Mask R-CNN 结构

Fig. 1 Mask R-CNN structure

算法对图像进行特征提取,根据具体目标需求与特点,设置 n 个不同的特征提取网络。针对手势关键点标定,选择 22 个残差网络[14],对输入图像进行处理。得以获得 22 个特征图,分别包含图像的不同深度信息。Mask R-CNN 利用 FPN[15]特征组合网络将不同深度的特征图进行重组,经过卷积、对位求和、上采样、池化等基本操作进行图像的重新生成,其中包含不同深度的特征信息。模型整体结构如图 2 所示。

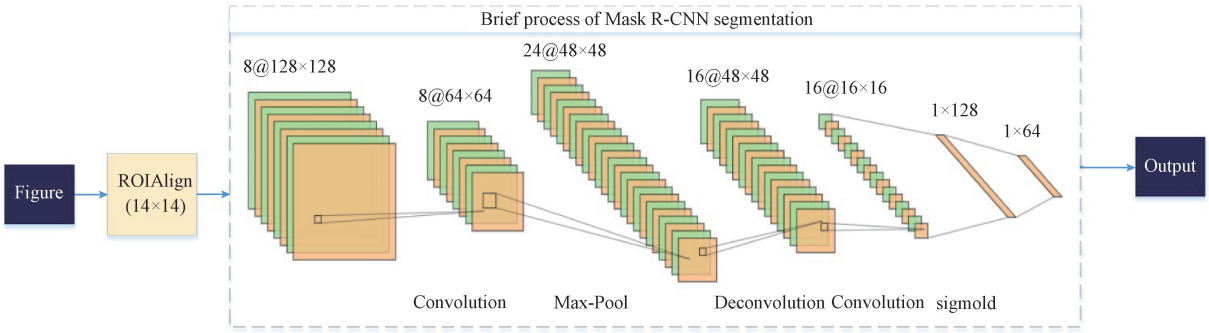


图 2　模型结构

Fig. 2　Model structure diagram

本方法采用 Anchor[16] 来完成区域提交功能。该部分组成模型结构中的区域提交网络。通过图像特征值大小,计算出能表示物体在图像中位置的多个候选框,并对结果进行修正。

对 x 个特征图中的每一个特征向量,进行回归计算。并将结果进行整合获得一个 n 维的向量,用以描述 n 个 Anchor 的修正值。每个修正值包括 Δx、Δy、Δh、Δw、p 五个量。分别代表新生成的候选框(box)与原始 box 的坐标、长宽修正值,以及前后景置信度。具体修正计算如下:

$$\begin{cases} x = (1 + \Delta x) \cdot x \\ y = (1 + \Delta y) \cdot y \\ w = \exp(\Delta w) \cdot w \\ h = \exp(\Delta h) \cdot h \end{cases} \tag{1}$$

式中:x、y、w、h 分别代表 Anchor 的中心横纵坐标,宽和高。经过 Anchor 修正后会产生大量的候选框,此时利用前后景置信度 p,通过非极大值抑制可得到较为精确的 box。

区别于原有办法从原图裁剪出相应区域并进行分割,本方法直接从特征图,利用 ROIAlign 算法[17] 直接裁剪出候选框对应的特征,加以双线性插值和池化,从而将特征图变换为统一的尺寸。采用 Softmax 层和全连接层的固定搭配,实现每个候选框与区域同一尺寸的特征的一一对应。并将结果作为头部功能性网络的输入参与后续计算。为了防止出现重复框选或选框过大造成的目标不明确问题,需要在将结果输入头部功能性网络之前,进行二次修正。即利用式(1)对当前结果进行计算,获得用以描述 n 个 Anchor 的修正值向量。其中,候选框各类别的形状的前后景置信度由 28×28 输出中的各点表示。最后,用 0.5 作为置信度阈值获取物体形状掩码,并经过一次全连接。最终可以获得目标的实例分割。本模型架构的实际效果如图 3 所示(以人为分割目标)。

1.2　骨骼点检测匹配与标定模型

人手作为一个小范围目标,很容易出现误识别问题。

图 3　Mask R-CNN 实例分割效果

Fig. 3　Mask R-CNN example

因此还需要进一步强化分割与标定模型。Simon 等[18] 提出一种 2D/3D 手部关键点检测方法[19]。通过利用立体几何信息,以多视图作为监督信号源,生成一致的手部关键点标签,引导训练手部关键点检测器。该方法通过弱监督训练[20],在训练数据上,只有少量标注数据,大量未标注的多视图数据,可以实时运行在单 RGB 图像上,其精度可与深度传感器方法媲美,并能够支持复杂对象 3D 无标记动作捕捉。本文选择基于该方法以及现有 31 视角手势骨骼点标定数据,对目标进行匹配与标注。

单视角图像容易由于遮挡等一系列原因,而导致部分点无法识别或错误识别。因在此多视角图像条件下,只需提取目标的部分未遮挡图像,即可根据视角的各自的位姿构建三角,得到具体 3D 位置信息。将所得到的点位置重投影到每一幅不同视角的 2D 图像,再使用这些 2D 图像和关键点标注训练检测模型网络。首先需要预设检测器,并根据已有数据对检测器进行预训练。

$$d(X) = \{(x_i, c_i) \text{for } i \in [1 \cdots I] \tag{2}$$

式中:d 表示检测器;X 为输入图像;x_i 与 c_i 分别代表预测关键点坐标以及置信度;I 表示预测点个数。根据真实数据对检测器进行训练后得到检测器 d_0,此时可以用该预训练检测器对未标注或误标数据进行训练。

$$t_{0,i} = \{F(t_{0,in}, t_{0,im})n, m \in [0 \cdots 31], i \in [0 \cdots 22]\} \tag{3}$$

$$T_0 = t_{0,1} + t_{0,2} + \cdots + t_{0,22} \tag{4}$$

式中:$t_{0,i}$ 代表第 1 组第 i 个骨骼点的真实数据;$t_{0,in}$ 与 $t_{0,im}$ 表示 31 个视角图像中目标清晰的两组。T_0 代表第 1

组 22 个关键点的真实数据集。

$$train(T_0) \rightarrow d_0$$
$$d_0 \rightarrow train(T_1)$$
$$train(T_0 \cup T_1) \rightarrow d_1 \tag{5}$$

式中：d_0 代表用第 1 组数据训练的检测器，对未标定数据进行预测标记，即 T_1。为避免新预测标定数据集存在与原始真实数据集的重复，需要进行额外监督处理，即进行二轮检测器训练。经过 n 次迭代，得到较为精确的手部关键点检测模型 d_n。

获得检测器后，通过 DNN 提取手势骨骼点识别模型权重。将图像转为 blob、forward 函数实现网络推断。并利用已训练的、效果较优的检测器，进一步获得手势关键点。该处理结构如图 4 所示。

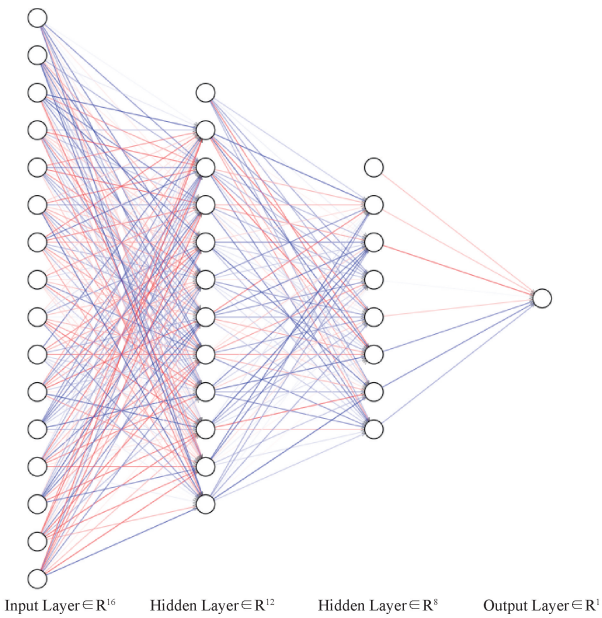


图 4 DNN 前向传播算法数学原理简图

Fig. 4 DNN forward propagation algorithm

通过网络计算可以获得手部 21 个关键点矩阵，分别代表预设定特定关键点的最大概率位置热图。调用 minmaxLoc 函数找到精确位置，即可实现对原始图像的标定。

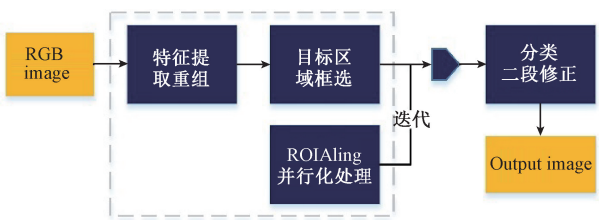


图 5 Mask R-CNN 骨骼点匹配标定流程简图

Fig. 5 Skeleton point matching calibration flow chart

2 模型改进与优化

由于手部处在高维运动空间，且待识别关键点间的曼哈顿距离[21]较小，骨骼点标注经常出现失真与跳变，造成了识别错误[22]。为了避免以上情况的出现，且降低时间损耗与计算成本，本方法在所构建的基于 Mask R-CNN 的模型后，加入 Savitzky-Golay 滤波器。通过对关键点的平滑处理与重新标定，大大提高准确性与稳定性。

Savitzky-Golay 滤波器[23]是一种在时域内，基于局域多项式最小二乘法拟合的滤波方法。能够在滤除噪声的同时可以确保信号的形状、宽度不变。常见数据滤波方法有平均移动法[24]（简单移动平均法、加权移动平均法）、指数滑动[25]等。但由于手部运动所带来的骨骼点坐标变化是没有规律的。因此为能最大程度达到保证原骨骼点标定正确，并达到防抖动与跳变的效果，本模型选用 Savitzky-Golay 滤波器进行优化处理。

首先将捕捉到的单帧图像存入数组，窗口长度设置为 n（正奇整数，本模型中取 19），每一个长度中的数据作为一个区间，记为 X 集合。

$$X = \{x_{n-m} + x_{n-m+1} + x_{n-m+2} + \cdots + x_n + \cdots + x_{n+m-1} + x_{n-m}\} \tag{6}$$

式(6)实现了将 X 从数据点的拟合转化为多项式拟合值的集合。对滤波窗口 $n(n = 2m + 1)$，采用 $k - 1$ 次多项式对窗口内的数据点进行拟合：

$$y = a_0 + a_1 x + a_2 x + a_3 x^2 + \cdots + a_{k-1} x^{k-1} \tag{7}$$

此后利用 n 个方程组成的 k 元线性方程组，通过最小二乘法拟合确定参数 σ：

$$\begin{pmatrix} y_{-m} \\ y_{-m+1} \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} 1 & -m & \cdots & (-m)^{k-1} \\ 1 & -m+1 & \cdots & (-m+1)^{k-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & m & \cdots & m^{k-1} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_{k-1} \end{pmatrix} + \begin{pmatrix} e_{-m} \\ e_{-m+1} \\ \vdots \\ e_m \end{pmatrix} \tag{8}$$

矩阵表示记为：

$$\boldsymbol{Y}_{(2m+1)\times 1} = \boldsymbol{X}_{(2m+1)\times k} \cdot \boldsymbol{A}_{k\times 1} + \boldsymbol{E}_{(2m+1)\times 1} \tag{9}$$

参数 σ 的最小二乘解及 $\boldsymbol{Y}$ 的模型滤波值（预测值）$\dot{Y}$：

$$\dot{\sigma} = (\boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{Y} \tag{10}$$

$$\dot{\boldsymbol{Y}} = \boldsymbol{X} \cdot \sigma = \boldsymbol{X} \cdot (\boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{Y} \tag{11}$$

进而对 $\boldsymbol{X} \cdot (\boldsymbol{X}^{\mathrm{T}} \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}^{\mathrm{T}}$ 进行求解，通过输入二位数组，并且每行采取最近邻补齐。对每一行进行

Savitzky-Golay 滤波,即可得到平滑后的新的骨骼点坐标数据,实现对关键点的精确绘制。

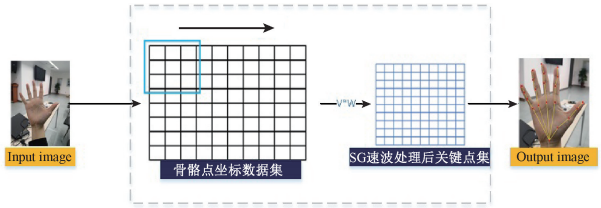


图 6　SG 滤波防抖平滑处理简图

Fig. 6　SG filter processing diagram

3　实验结果及分析

3.1　骨骼点检测与标定实验

本文采用基于 Mask R-CNN 与 SG 滤波的骨骼点识别标定算法,利用相关点之间的曼哈顿距离对骨骼点分类进行了进一步细化,组成手势关键点提取系统。

为精确标定手语手势骨骼点,采集了 18 组手势动作,共 90 组视频序列作为数据集进行了训练。每组包含 5 种场景,3 种光源条件(正常光、强光、弱光),两种状态(手部特写及全景,全景即图像中手部非最大连通区域)。此外,引入 DEVISIGN 手语数据集,扩充样本集合。经过 873 次迭代后获得测试模型。

针对模型在四卡服务器上进行了关键点标定实验,共设定识别组,运算速度,精度 3 个评估指标。随机抽取 50 组动作视频序列组成集合 Y, 作为实验数据并逐帧进行处理。首先人工标定关键点位置区间,作为关键点运动范围。模型计算获得标定点坐标序列后,与人工标定结果进行比对,若在人工设定范围内,则为标定成功。计算识别正确标定点占总数的百分比。百分比平均值即为该算法精度值。并将 50 组中含有识别失败点的视频序列归为集合 W。针对识别组指标,随机抽取集合 Y 中的 30 组数据结果。若某序列标定精度值大于 80%,则表示该组识别成功。此外,在模型算法中引入 time 评价,用以计算程序运行平均消耗。

相比于传统的手部关键点提取算法,以及未进行骨骼点进一步分类细化与 SG 滤波的模型,本模型大大提高了骨骼点识别精度。针对上述 3 种模型,在集合 Y 上分别进行了评估,结果如表 1 所示。

表 1　模型性能评估

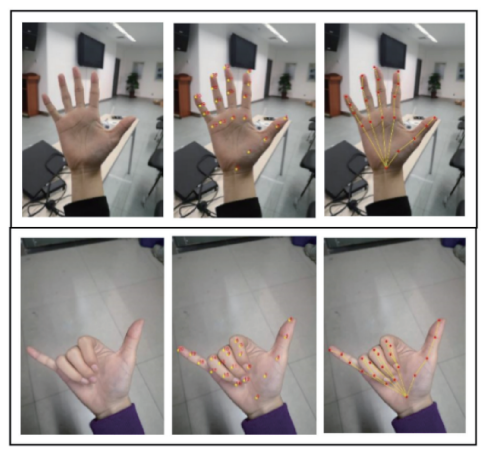
Table 1　Model performance evaluation

算法	可识别组	平均用时/ms	AP
传统算法	12/30	867	49.63
Mask R-CNN(caffe 2)	25/30	203	82.97
本文算法	28/30	142	93.48

此外,为判断本方法在不同环境下的误差,从而进行更全面的效果评估,还进行了误差计算。

通过对 50 组中 3 种光照条件的不同结果进行统计,每组的错误标定点占总数的百分比及为误差。结果表明该方法的平均误差(两种状态下),正常光条件下小于 5%,强光与弱光条件下误差最大分别为 4.73% 与 9.51%。部分实验结果如图 6 所示。图 7(a)所示为全景正常光状态下的手部骨骼点识别结果,分别为原始图像、关键点标注图以及关节连接图 3 部分。图 7(b)所示为正常光与弱光条件下的手部特写图实验结果。可以看到关节点标注误差控制在单动作 2~3 个关键点。

(a) 全景正常光状态下的手部骨骼点识别结果
(a) Experimental results of close-up pictures of hands under normal light and low light conditions



(b) 正常光与弱光条件下的手部特写图实验结果
(b) Experimental results of close-up pictures of hands under normal light and low light conditions

图 7　实验结果

Fig. 7　Experimental results

由实验结果表明,本文所构建的骨骼点标定模型,在计算速度,平均精度,及可识别组别 3 个指标上,都强于其他算法。本方法实现了精度与运算速度的提升。但在识别组数的提升上优势不明显。

3.2 数据平滑滤波实验

手部骨骼点识别由于存在高维失真以及关键点跳变抖动的问题,选用滤波的方法对模型进行了优化,并针对不同滤波器的进行了效果对比试验。

常用数据平滑滤波器有滑动平均法(简单移动平均法、加权移动平均法)、指数滑动(1 次、2 次、3 次)等等,本文针对 3 大类 6 种常用数据平滑滤波进行了对比测试。

该实验选用 20 组手势动作(8 组识别正确,以及骨骼点检测匹配与标定实验的集合 W 中的 12 组误识别或未识别手势数据集)进行平滑测试。参考光流法[18]中稠密光流与稀疏光流对目标像素点移动的捕捉,在实验中,对每组手势数据集的原始数据轨迹,以及利用不同滤波器平滑防抖处理后的关键点运动范围,进行了绘制,轨迹绘制效果如图 8 所示。

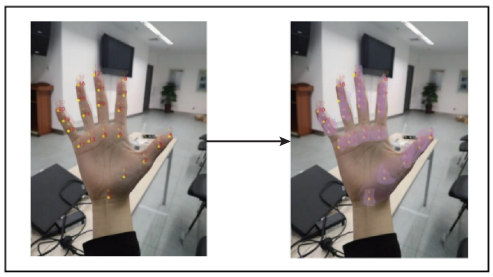


图 8 抖动范围标定实验

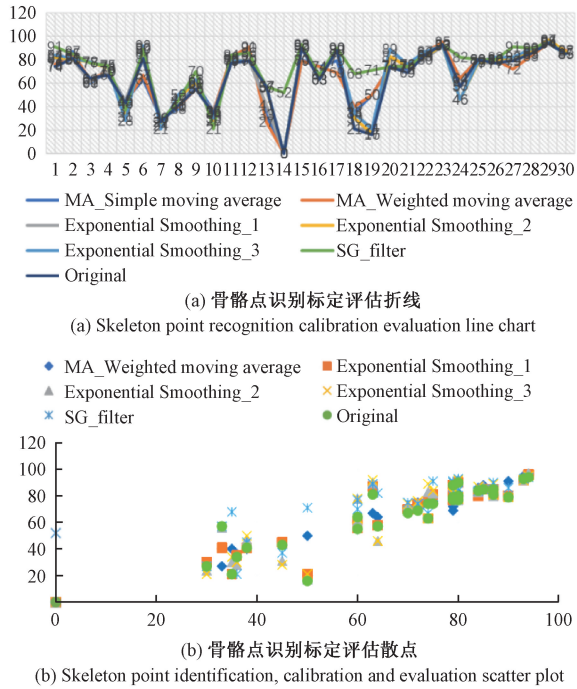
Fig. 8 Calibration experiment of jitter range

从而可以得到关键点原始抖动范围与面积,基于该指标可实现防抖动。融合误识别以及未识别点指标权重,对不同滤波器在模型中的优化效果进行评估。评估结果如图 9 所示。

从图 9 可以看出,与原始数据相比,SG 滤波器对于手势关键点的防抖平滑优化效果相对较好;滑动平均两种方法简单平均权重一致,精度无法达到基本要求;加权滑动平均则由于是平均值,预测值总是停留在过去的水平上,而无法预计会导致将来更高或更低的波动,优化效果不明显;指数滑动法虽然相对滑动平均效果较优。但由于手势运动无规律,该方法所预测的处理后最优解,将导致指数预测滞后于实际需求从而出现较多失真。SG 滤波器对手部骨骼点的防跳变效果以及稳定性,包括普适性明显优于前两者,总体使骨骼点识别匹配模型精度与鲁棒性,得到了较大提升。

4 结论及展望

本文在 Mask R-CNN 的基础上,提出一种手势关键点 Mask R-CNN 模型与 Savitzky-Golay 滤波的提取方法。



(a) 骨骼点识别标定评估折线

(a) Skeleton point recognition calibration evaluation line chart

(b) 骨骼点识别标定评估散点

(b) Skeleton point identification, calibration and evaluation scatter plot

图 9 手势骨骼点识别标定评估结果

Fig. 9 Evaluation results of gesture skeletal point recognition calibration

基于 Mask R-CNN 模型处理 RGB 三通道图像,对每张图进行物体识别和分割,并利用神经网络特征对视频流进行目标匹配,得到手部 21 个关键点。进而利用 Savitzky-Golay 滤波对数据进行平滑防跳变处理,得到精确稳定的手势关键点提取重建结果。该模型在不同光照条件下平均精度最高可达 93.48%;在四卡服务器运行条件下识别速度达到 142 ms。实验测试结果表明,本文方法能够最大程度摒除环境干扰,精准提取关键点,与传统方法及单一的 Mask R-CNN 提取相比[26],在精度与鲁棒性上都明显提高。

参考文献

[1] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. Proceedings of the Conference and Workshop on Neural Information Processing Systems (NIPS), 2014.

[2] ZIMMERMANN C, BROX T. Learning to estimate 3D hand pose from single RGB images[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[3] MOLCHANOV P, YANG X D, GUPTA S, et al. Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks [C]. Proceedings of the IEEE Conference on Computer Vision

and Pattern Recognition (CVPR), 2016.

[4] KRISHNAMURTHY J, MITCHELL T M. Weakly supervised training of semantic parsers [C]. Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (ACL),2012.

[5] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017.

[6] HU T, WANG W, LU T. Hand pose estimation with attention-and-sequence network [C]. Proceeding of the Pacific Rim Conference on Multimedia,2018.

[7] YANG S, LIU J, LU S, et al. Collaborative learning of gesture recognition and 3D hand pose estimation with multi-order feature analysis [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2020.

[8] HE K, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV),2017.

[9] XIE S, GIRSHICK R, DOLL'AR P, et al. Aggregated residual transformations for deep neural networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[10] HUANG J, SHAO X, WECHSLER H. Face pose discrimination using support vector machines (SVM) [C]. Proceedings of the International Conference on Pattern Recognition (ICPR),1998.

[11] GIRSHICK R. Fast R-CNN [C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.

[12] 颜雪军,赵春霞,袁夏. 2DPCA-SIFT:一种有效的局部特征描述方法[J]. 自动化学报, 2014, 40(4): 675-682.
YAN X J, ZHAO CH X, YUAN X. DPCA-SIFT: An efficient local feature descriptor [J]. Acta Automatica Sinica, 2014, 40(4):675-682.

[13] DE LA GORCE M, FLEET D J, PARAGIOS N. Model-based 3D hand pose estimation from video [J]. Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2011, 33(9): 1793-1805.

[14] TOMPSON J, STEIN M, LECUN Y, et al. Real-time continuous pose recovery of human hands using convolutional networks [J]. ACM Transactions on Graphics (ACM TOG), 2014, 33(5): 1-10.

[15] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[16] DAI J, HE K, SUN J. Convolutional feature masking for joint object and stuff segmentation [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2015.

[17] GIRSHICK R, IANDOLA F, DARRELL T, et al. Deformable part models are convolutional neural networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[18] SIMON T, JOO H, MATTHEWS I, et al. Hand keypoint detection in single images using multiview bootstrapping [C]. Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017:1145-1153.

[19] PRESS W H, TEUKOLSKY S A. Savitzky-Golay smoothing filters [J]. Computers in Physics, 1990, 4(6):869-872.

[20] PINHEIRO P O, LIN T Y, COLLOBERT R, et al. Learning to refine object segments [C]. Proceedings of the European Conference on Computer Vision (ECCV), 2016.

[21] TANG D, CHANG H J, TEJANI A, et al. Latent regression forest: Structured estimation of 3D articulated hand posture [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2014.

[22] HUANG J, RATHOD V, SUN C, et al. Speed/accuracy trade-offs for modern convolutional object detectors [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[23] WANG R Y, POPOVI'C J. Real-time hand-tracking with a color glove[J]. ACM Transactions on Graphics (ACM TOG). 2009, 28(3): 1-8.

[24] YE Q, YUAN S, KIM T K. Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2016.

[25] SUN D Q, ROTH S, BLACK M J. Secrets of optical flow estimation and their principles [C]. Proceeding of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), 2010.

[26] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: New benchmark and state of the art analysis[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2014.

作者简介

王婧瑶,北京工业大学,人工智能与自动化学院学生,主要研究方向为计算机视觉、机器学习、自主机器人、三维重建。
E-mail:924944184@qq.com

Wang Jingyao is a B. Sc. candidate at Beijing University of Technology. Her main research interests include computer vision, machine learning, autonomous robot and 3D reconstruction.

王红军(通信作者),2005 年于北京理工大学获得博士学位,现为北京信息科技大学教授,主要研究方向为高端装备智能感知与控制、故障诊断与维护。
E-mail:wanghj86@163.com

Wang Hongjun (Corresponding author) received her Ph. D. degree from Beijing Institute of Technology in 2005. Now she is a professor at Beijing Information Science and Technology University. Her main research interests include high-end equipment intelligent perception and control, fault diagnosis and maintenance.