三维手势姿态估计算法研究

方桂安，刘梦莎，张焯霖，刘玥，唐迅

摘 要—本次作业的目标是实现基于深度图像的手部姿态估计，我们在 paperwithcode 的帮助下，对近些年来该领域的研究工作进行总结和梳理，阅读并复现榜单上各种各样的方法，也在老师提供的数据集上进行了我们自己的实验与改进，采用主流的三维平均关节误差作为评价指标评估，最终取得了很好的分数。

关键词—深度图像，手部姿态估计，AWR

I. 概述

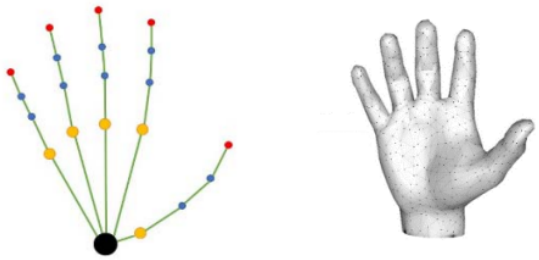
基于视觉的裸手交互一直是人机交互、虚拟现实和增强现实领域的研究热点, 其应用十分广泛, 如在命令交互、手语识别和真实感抓取等多个方向中都需要准确地进行手部姿态估计。

现有的手部姿态估计方法可以分为有标记的手部姿态估计和无标记的手部姿态估计。前者主要通过佩戴专用设备获取手部关节位置或关节旋转角度, 如数据手套和彩色标记手套等, 或者在场景中布置复杂的多摄像头采集设备, 如光学运动捕捉系统等。从用户体验角度, 最理想的是基于视觉的无标记方法, 即手部无须佩戴任何设备就能准确进行手部姿态估计。传统方法主要是通过彩色光学相机采集手部运动图像, 再运用计算机视觉相关算法估计手部姿态, 其准确性不高且鲁棒性很难达到实用需求。其主要原因是光学相机采集图像的过程中丢失了深度信息, 使从彩色图像恢复手部姿态面临二义性问题; 同时, 手运动速度较快且自由度高, 而不同人的手部肤色多变, 极大地增加了解决问题的难度; 而且, 图像背景环境复杂, 图像质量容易受到环境光的影响。

II. 手势估计问题介绍

基于深度图像的手部姿态估计的问题定义为从单幅或连续深度图像中恢复手部关键点的位置 (二维的图像坐标或三维的空间位置) 或手的自由度参数。输入的是包含手部的深度图像, 输出的是二维或三维手关节的位置, 或是若干个自由度参数。

图 1a 所示为一个常用的 26 自由度手部动力学骨骼模型, 其中不同的关节具有不同的自由度 (根关节点包换全局位移和旋转角度)。图 1b 所示为骨骼模型经过前向动力学和线性混合蒙皮后得到的手部几何面片模型。在获得各个关节的旋转角度后, 结合手的前向动力学和给定的手部骨骼铰链结构可以重建出手的关节位置和表面细节等信息。另外, 手的自由度维度远小于手的关节点坐标的维度, 一些方法利用该特性来降低估计问题的复杂度。



a. 手部骨骼模型　　b. 手部面片模型

Fig. 1. 手部骨骼模型与其对应的面片模型

根据输入的深度图像是单幅深度图像还是连续的深度图像序列, 可以将现有的方法分为 2 种:

1) 直接从单幅深度图像中恢复手部姿态;
2) 从连续深度图像序列中估计出连续的手部姿态。

它们的区别在于, 后者可以利用深度图像序列的时序信息, 该方法有时也被称为手部追踪 (hand tracking)。

根据深度相机和手的位置关系可以将现有的方法分为 2 种:

1) **正向视角**, 主要应用于基于姿态控制的系统中, 在这种视角下手掌正对深度相机, 手在图像中所占面积较小且手运动范围较大;
2) **自我视角**, 主要用于虚拟现实和增强现实场景, 在该视角下手距离深度相机较近, 手腕离摄像机更

方桂安，20354027，(e-mail: fanggan@mail2.sysu.edu.cn)。
刘梦莎，20354091，(e-mail: liumsh6@mail2.sysu.edu.cn)。
张焯霖，20354156，(e-mail: zhangzhlin8@mail2.sysu.edu.cn)。
唐　迅，20354121，(e-mail: tangx66@mail2.sysu.edu.cn)。
刘　玥，20354229，(e-mail: liuy2236@mail2.sysu.edu.cn)。

近, 手在深度图像中面积较大且自遮挡严重 (指尖经常被遮挡或移出图像之外)。

图 5所示为 2 种不同视角下的手的彩色图像和对应的深度图像。

a. 彩色图像　　b. 深度图像

Fig. 2. **自我视角和正向视角下的手部图像**

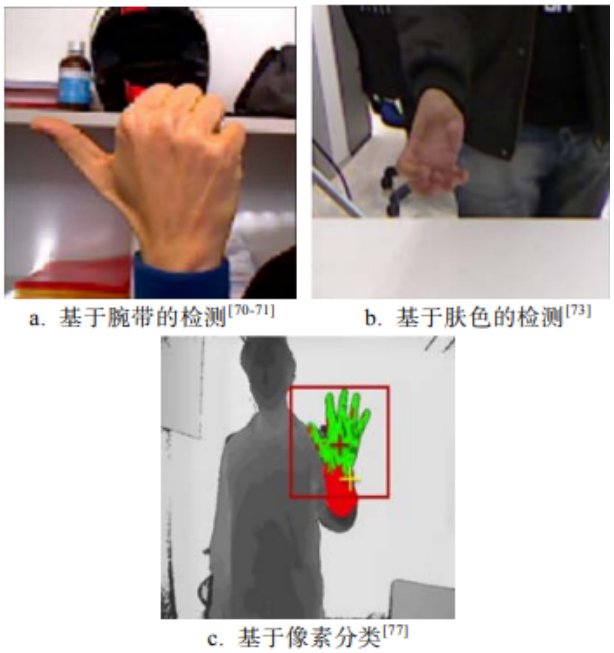
目前为止, 基于深度图像的手部姿态估计仍然是一个尚未被很好地解决的问题, 其主要面临以下挑战。

(1) **手的自相似性和自遮挡问题**。人的各个手指的外形相似度较高, 这增加了辨识的难度. 手的自由度较高, 铰链物体的高自由度带来的一个最直接的问题就是手指之间的自遮挡严重, 这些严重的自遮挡可能导致手指的不同部分在图像中重叠。手部的自遮挡还包括 2 只手交互时相互之间发生的遮挡。这些情况有时也被称为不完整数据的问题。此外, 在虚拟现实和增强现实场景中还面临其他需要解决不完整数据问题。例如, 手经常只有一部分在视野中, 或者经常移出视野或被其他物体遮挡了一部分。

(2) **深度相机硬件性能有限**。受深度传感相机成像原理限制, 现阶段的大多数深度相机在实时采集速度 (30 帧/s) 情况下, 最高分辨率一般为 480 像素, 部分双目视觉相机分辨率能够达到 720 像素, 但是成像质量较差且帧率低 (仅有 7 8 帧/s)。人的手具有快速移动能力, 手腕的移动速度甚至可以达到 5 m/s. 在采集速度仅有 30 帧/s 的情况下, 意味着前后 2 帧之间手的位置可能发生很大变化, 导致输出存在着空间上的不连贯现象 (该现象也称为抖动)。值得注意的是, 手的快速运动还会导致深度图像产生严重的运动模糊和特有的边缘噪声, 这些复杂的噪声也增加了估计问题的难度。当前深度相机的工作距离比彩色光学相机小, 这也是一个问题, 具体体现为常用的深度相机 (如 Intel RealSense) 的有效范围仅为 0.2-0.6 m, 绝大多数深度相机距离超过 1.0 m 后获取的深度图像噪声严重, 甚至无法识别手部轮廓。另外, 深度相机的视场相比光学相机狭窄。

III. 主流工作

大多基于深度图像手部姿态估计方法都默认手部区域已经检测并分割出来。这些方法大多数通过基于最近物体阈值的原理实现手部定位, 即根据深度距离阈值分割出手和背景。还有一些基于颜色的方法, 如 Tagliasacchi 等人使用了基于腕带的手部检测方法 (图 3(a)), 检测的手腕方向作为额外的约束项 (通过手腕的朝向可以额外计算出手的旋转信息) 嵌入姿态优化方程中。Choi 等使用卷积神经网络同时进行手的检测和分割, 直接使用整个深度图像作为输入, 同时输出手部检测框、手的中心位置和手部语义分割。Oikonomidis 等用基于肤色的手部检测算法分割手部区域, 图 3(b) 所示为基于肤色的手部区域分割结果。还有一些方法使用人体关节点跟踪器估计手的粗略位置。Taylor 等将手部检测视为一个手和背景的二值分类问题, 其分类结果如图 3(c) 所示。另外, 还出现了使用卷积神经网络 (convolutional neural networks, CNN) 进行手部检测的方法。



a. 基于腕带的检测[70-71]　　b. 基于肤色的检测[73]

c. 基于像素分类[77]

Fig. 3. **几种常见手部检测方法**

A. 基于模型驱动的方法

基于模型驱动的方法将手部姿态估计转化为点云配准的优化问题, 其试图最小化点云和一个预先设定好的参数化手部模型间的距离残差。由于该问题高度非凸且手部自由度维度高, 遍历整个搜索空间寻找最优解往往无法满足实时需求。从理论上讲, 现有方法不管采用全局最优还是局部最优, 本质上都是寻找某个姿态使目标函数在某处的残差最小。

现有模型驱动方法对优化方程的初始值非常敏感。因此, 很多方法采用启发式的姿态初始化和局部搜索相结合的方法求解次优解, 有的方法默认手一开始就处于屏幕中心, 这也常被认为一种隐式的初始化方法。另外, 一些方法用前一帧的结果作为下一帧的初始值, 这种方法往往由于累积的估计误差而极容易陷入追踪失败, 而手的高自由度导致模型驱动方法很难从追踪失败中恢复, 这也是基于模型驱动的方法面临的最大问题。最近的一些工作在采用了基于模型驱动方法的同时结合了深度学习, 这些方法归纳于混合驱动方法并在后续章节进行介绍。在介绍相关工作前, 首先对模型驱动方法涉及的一些关键技术进行介绍。

手的前向力学 (forward kinematic, FK) 描述了如何将手部自由度参数映射为手的骨骼模型和几何面片模型。给定手部自由度向量和预定义的手层次骨骼模型, 根据前向动力学可以计算出手各个关节点位置 (即骨骼模型), 再结合线性蒙皮算法, 可以进一步获得手部的几何网格模型。与此相反, 从手部关节点位置计算出手部自由度的方法称为逆向动力学 (inverse kinematic, IK), 其常用于模型驱动的离线优化过程中。PSO 是一种全局优化算法, 其通过在若干迭代次数内进化若干粒子实现寻找最优解。PSO 非常适合于求解非光滑目标函数, 常用于非凸问题寻找全局最优解。早期模型驱动方法基本都采用 PSO 估计手部姿态。

最近点迭代 (iterative closest point, ICP) 算法常用于三维点云配准问题。Fleishman 等提出了结合向运动学的 ICP 算法。Qian 等发现, ICP 能够快速达到局部最优, 而 PSO 能更有效地探索整个参数空间, 但存在过早收敛的问题, 因此他们提出了 ICP-PSO 方法, 即将 ICP 和 PSO 结合并搜索整个参数空间, 使其迅速收敛到某个局部区域得到最优解, 其在单个 CPU 上的速度达到 24 帧/s。

对于同样是采用 ICP 的基于模型驱动的方法, 根据每次迭代时计算梯度雅可比矩阵的区别, 可以分为数值求解方法和解析求解方法。由于手的自由度参数较多, 因此手的前向过程计算复杂高, 进而导致前者的计算复杂度远高于后者, 这也是 PSO 方法远慢于梯度下降法的主要原因之一。

Tkach 等提出了基于 ICP 的手部姿态优化方法, 其使用经典的 “渲染-计算” 框架, 并公开了其代码, 对后续工作影响很大。该方法使用深度和彩色图像作为输入, 并使用圆柱体表示手部模型, 进行优化求解。这个过程使用 ICP 获得三维点云和模型对应关系, 利用距离变换算法提取二维轮廓生成像素-模型对应关系, 使用梯度下降算法迭代计算进行优化, 其中使用了手部关节旋转约束、时序约束、手部运动空间约束和关节碰撞约束等。但该框架最大的瓶颈是, 在模型绘制时上下文的切换耗时很大。

除了传统的基于几何优化的模型驱动方法, 还有一些方法试图引入刚体动力学、高斯混合模型估计手部姿态。Melax 等采用基于物理模拟的方法对手部模型和深度点云的配准, 使用 17 个圆柱刚体模型表示手, 通过求解刚体动力学方程解决手部姿态估计问题。在求解过程中同时引入刚体碰撞检测等手部先验约束, 并在数据约束项中使用基于点到平面的三维数据约束, 使收敛过程更加平滑。Sridhar 等用高斯模型表示手 (即在手的关节等处用高斯模型代替手部几何面片模型), 将输入的点云和手部都表示为高斯分布, 最后用局部梯度下降的方法优化手部姿态。该算法在保持较高精度的同时, 仅需要 CPU 硬件的支持即可达到实时效果。

大多数基于模型驱动的方法采用固定大小的三维手模型匹配点云, 但不同人的手大小不同且关节长度比例也不同。手的自由度信息除了包含各关节的旋转角, 还应该包含各关节之间的距离, 用固定尺寸的三维手部模型进行优化, 会降低算法的准确性。因此, 在优化过程中, 考虑将手的长度作为变量很有必要。Khamis 等在优化过程中将手部骨骼的长度作为变量加入优化过程, 只需要使用 15 幅连续深度图像就能完成离线手部形状的标定。Remelli 等提出了离线手部标定算法, 使用一个球体模型表示三维手部模型, 将手的形状作为变量加入优化过程中进行全自由度优化 (即同时优化手的姿态和形状), 在优化目标方程中同时引入了点云-模型的误差和模型-点云的误差。图 4 所示为其多阶段优化的过程。

模型驱动的方法极度依赖于初始解的选择和优化方程的目标函数设计, 求解得到的姿态往往具有空间上

Fig. 4. Remelli 等联合手部形状进行优化的示意图

的连贯性且符合物理自然约束。因此, 近些年来基于模型驱动的方法侧重于研究不同的手部模型表示 (如采用可变尺寸的手部模型或探讨如何在模型精细化和计算速度之间均衡), 或者采用不同的先验约束项确保求解得到的手部姿态的结果更为真实。但是, 由于消费级深度相机采集的深度图像质量较差, 模型到点云的配准过程中易产生误差累积使追踪失败。仅仅依赖一些启发式的方法很难从追踪失败中恢复出来, 这也是基于模型驱动的方法最大的缺点。

B. 基于数据驱动的方法

基于数据驱动的方法从大量的训练数据中学习一个从深度图像到手部姿态的映射函数。与基于模型驱动的方法不同, 基于数据驱动的方法用带标记的手部深度图像训练分类器或回归模型。这些经过训练后的分类器或回归模型, 可以从输入的深度图像序列中恢复出对应的手部姿态, 直接处理单幅深度图像, 不需要基于模型驱动方法的初始化过程, 因而追踪过程不存在误差累积。例如, 手从相机视图中消失后再次进入, 系统能继续运行。基于数据驱动的方法的最显著的特点是必须依赖足够多的训练样本才能实现较好的泛化能力, 这里的“足够多”是指抽样的手部姿态能够均匀地覆盖整个手部运动空间。下面按照所用的分类器/回归模型种类, 论述一些典型的基于数据驱动的方法。

Keskin 等延续了 Kinect 人体骨架关节点追踪的“逐像素部件分割 + 区域聚类中心”的模式, 对输入的手部图像使用训练过的随机森林 (random forest, RF) 进行逐像素部件归属分类, 并使用均值漂移算法寻找类的中心, 然后将寻找到的位置向 Z 方向平移, 获得最终的关节点。该方法对于某些极端视角下的输入图像的泛化能力较差。Tang 等使用隐变量回归方法, 直接从深度图像中估计手部姿态, 其根据手的拓扑结构由粗到细层级地寻找隐节点对应拓扑结构的重心位置, 最终找到手的各个节点位置。

Sun 等将传统的二维图像的级联姿态回归拓展到三维手部姿态估计上, 优化了以往基于 RF 方法使用的二维像素差异特征, 同时针对手的特殊拓扑结构提出了三维层次级联回归的方法。具体做法如下:

1) 对手进行归一化以增加旋转不变性;
2) 对归一化后的手部关节位置提取三维姿态索引特征;
3) 基于这些三维姿态索引特征进行层级级联回归。

与之前的级联算法不同的是, 该算法先估计手掌的位置, 然后分别估计各个手指的位置。这是因为手掌的位置相对于手指更加稳定, 而且手掌位置的变化更会显著地影响手指的估计, 且手指之间的估计是互相独立的。Sun 等分别在 ICVL 和 MSRA15 这 2 个数据集上进行了验证, 由于采用了层级学习残差和手掌/手指分开回归的策略, 与隐变量回归 RF 相比, 其在指尖的估计精度上有提高。该工作是目前基于 RF 方法中准确度和速度最高的。基于 RF 的方法大多集中于 2015 年以前, 根据其公开结果, 这些方法都能够满足实时的追踪需求。但是由于在这些方法中选取的都是局部梯度特征, 该特征只能描述某个点附近的三维局部纹理信息, 而深度图像的噪声会随着手的运动速度剧烈增加。如果输入的图像噪声过大, 会导致 RF 方法的准确性迅速下降。CNN 相对传统机器学习方法在准确性和鲁棒性等方面有很大优势。Tompson 等用多分辨率输入的 CNN 将原图根据平均深度归一化, 分别降采样到 3 种分辨率, 再分别输入网络生成的关节点分布热图。这种多分辨率输入能够使其网络分别学习到全局和局部特征, 同时, 该策略也对后续的很多工作具有借鉴意义。

2015 年开始, 出现了大量了基于 CNN 手部姿态估计的方法, 这些工作采用不同的网络结构提取手部特征。Oberweger 等用 CNN 回归手部姿态, 其使用多尺寸和多阶段的 CNN 直接做姿态的回归, 第 1 阶段使用网络回归生成粗略的手部关节位置, 第 2 阶段回归生成第 1 阶段关节位置和样本真值之间的残差 (即采用多阶段级联的策略)。该方法的一个贡献在于提出基于主成分分析的姿态先验层, 并基于此提高了预测精度, 但实际上普遍认为这是隐式地进行参数降维, 并没有真正利用手的先验知识。

有些方法尝试将手的先验知识显式地加入神经网络中。Zhou 等将手的前向过程嵌入 CNN 中, 提出了一个无参数的手模型层 (hand model layer), 该层将手部自由度通过可微的前向动力学函数映射为手部关节点坐标, 同时使用关节旋转的约束作为新增的损失函数项;

但是由于其选取的特征提取网络比较简陋, 因此准确度比同时期其他工作提升有限。

除了传统的二维图像作为神经网络的输入外, 还有尝试采用其他方式表示深度图像的工作。Ge 等首先将输入的深度图像重新投影到 Y 平面和 Z 平面上, 通过多个卷积层各自生成对应的热力图, 然后进行多视角融合输出最终的关节位置。与传统的二维图像相比, 这些三维表示能够更好地描述深度图像的特征, 并学习到多个视角下的点云特征。

除了采用不同类型的图像输入外, 使用基于无序点云的网络结构也是一大研究热点。Ge 等用基于点云的网络进行手部估计, 比传统 CNN 提高了准确性。其直接使用 PointNet 作为特征提取模块进行姿态估计, 一个创新点在于提出基于包围盒的数据增强方式和指尖增强网络用于提高指尖关节的预测精度, 这种增强网络本质上是一种级联式的学习方式。其提出的基于包围盒的数据增强方式能在一定程度上解决深度点云噪声过大的问题。Chang 等使用基于“三维体素输入-三维体素热图输出”的方式进行手部姿态估计, 其网络使用体素化的深度点云作为输入, 逐体素地对手部关节点位置进行最大似然估计。

一些工作也试图将自/半监督学习引入手部姿态估计中。Wan 等提出了一个自监督的框架解决手部态中训练样本难以获取的问题。其输入图像首先经过一个姿态估计网络 (在合成数据上训练而成) 输出手的关节位置, 并通过这些关节位置生成三维的手部球体模型并投影到指定平面上, 然后使用这个投影出来的合成深度图像和输入的深度图像进行模型匹配, 并将其作为整个网络的损失。该方法能够很好地解决现阶段关于手的训练样本不够的问题。

2018 年后, 已经没有太多针对单手的基于深度图像的姿态估计工作出现在三大计算机视觉顶会, 研究热点开始转移到估计抓取物体时的手部姿态, 其中具有代表性的工作是 Doosti 等提出的 HOPE-Net, 该工作通过训练过的神经网络同时估计手部姿态和物体的姿态。

近年来, 基于数据驱动的方法成为整个领域的主流, 相关研究也大多集中在这个方向, 而它们在公开数据集上的表现也证明了其准确度比基于模型驱动的方法有了很大的提升。根据网络输出方式的不同可以将这些工作分为检测方法和回归方法, 前者使用关键点热图回归的方式恢复出手的关键点位置的最大似然概率, 后者则直接通过全连接层输出归一化后的关键点坐标。前者的准确度取决于热图的分辨率, 由于逐像素的计算耗时较多, 后者一般在特征提取网络后直接用全连接层输出关键点位置信息, 缺点在于准确度不如检测方法。从网络的输入角度可以分为使用二维深度图像、三维体素、无序点云和图卷积等。一系列工作也证明了与单纯的二维网络相比, 三维数据结构能够更好地表示手部的几何信息。

同时, 基于数据驱动的方法也仍然面临着很多问题。首先, 大多数基于数据驱动的方法只使用单幅图像作为输入, 无法有效地利用时序信息, 预测的结果存在着空间上的不连贯性。其次, 这些网络模型并没有真正意义上利用手的各种先验信息, 使其预测的结果往往不符合物理约束。这些方法还有一个最明显的问题在于, 其网络对于输入数据非常敏感, 而剧烈运动导致的大量噪声可能降低整个网络模型的准确度。

C. 混合方法

基于模型驱动的方法和基于数据驱动的方法在许多方面存在互补性。基于模型驱动的方法需要姿态初始化才能进行追踪, 而基于数据驱动的方法往往无法利用手的先验信息而使结果缺乏空间连贯性。因此, 一些工作提出的混合方法将上述 2 种方法结合在一起, 用于解决手部姿态估计问题。其使用一个训练过的分类器或回归模型对手部姿态进行初始化, 这里的初始化的用途可以是寻找逐像素的点云-模型的对应关系, 也可以是利用数据驱动的方法估计优化方程的初始近似姿态值。通过初始化得到方程初始值或语义上的点云模型对应关系的优化方程, 能够更快速、准确地求解最终结果。

Krejov 等将基于 RF 逐像素分割的方法和基于物理的模型拟合的方法相结合, 先利用 RF 进行手的区域分割, 再利用基于刚体动力学的方法进行优化, 这种混合策略能从由复杂视角导致的手部姿态丢失的情况下迅速恢复。Sridhar 等也采用这种“区域分割 + 模型拟合”的策略进行手部姿态估计, 但是其使用了高斯混合模型代替前者基于物理的优化方法。

Sharp 等提出了一个实用的手部追踪系统, 声称其具有准确 (对不同手势都有准确的估计结果)、鲁棒 (使用初始化器, 使系统能从各种追踪失败中迅速恢复) 和自由 (该系统只需要一个 Kinect v2 摄像头且对于手的位置和角度没有特殊限制)3 大特点。该系统包含了重初始化模块和模型拟合模块。重初始化模块使用了 2 层分类器的方法, 第 1 层用于预计手的全局旋转, 第 2 层

用于预测具体的手部姿态。与之前使用 RF 直接回归手部姿态不同，这里的第 2 层分类器只是产生姿态的分布，以进行一系列不同的姿态的估计。模型拟合模块将粒子算法和基因演化算法相结合，通过定义黄金能量函数为对应位置深度的截断距离的和生成多个姿态，使用基因算法的策略对于一些能量比较高的粒子进行随机变换，最终经过多次迭代，并选取最低能量的粒子所代表的姿态作为最终结果。

Taylor 等延续了渲染-计算的框架，但是使用一个手势索引 RF 代替了原来的 2 层手部姿态初始化 RF，又用基于梯度下降的优化方程代替了原有的基于 PSO 和基因演化算法的模型匹配。其中，使用融合了时序信息和手部姿态检索 RF 的输出作为当前的初始化姿态，从这个初始化姿态出发，结合多种手部先验约束（如碰撞、时序、关节旋转等）进行手部姿态优化。在后续工作中，Taylor 等将该算法拓展到交叉的双手上，使用一个手部分割网络区分出左右手的像素归属，再用基于梯度下降的方法优化手部姿态。在优化过程中借鉴了现有的 SLAM 系统中的多种思路，如用卡尔曼滤波平滑前后帧的姿态。

混合方法在某种意义上结合基于数据驱动的方法和基于模型驱动的方法的优势，使用数据驱动作为姿态初始化工具可以很快地从追踪失败中恢复，而使用优化方程进行点云-模型配准可以使输出结果在满足空间上的连贯性的同时，保持物理约束。当遇到一些噪声过大的图像时，能够使用时序信息和运动空间约束平滑当前的结果。这种组合多种方法的策略在一定意义上使混合方法的输出结果非常平滑，这对于整个人机交互系统非常重要，用户在使用裸手进行交互时不会感觉到迟滞感，也能够及时地体会到视觉反馈。

IV. 模型说明：AWR

为解决手势姿态估计问题，北京大学团队提出了用于 3D 手部姿态估计的自适应加权回归方法（AWR），它兼具了基于回归和基于检测方法的优点，利用可导的信息聚合的方法统一了关节点的稠密特征和关节点坐标回归，梯度能从关节点坐标回传到稠密特征，使网络能够端到端训练。另外，自适应的权重图使得网络能应对关节点附近深度值缺失、关节点被遮挡或者指间自相似性等复杂问题，增强网络的鲁棒性。大量实验证明了 AWR 在不同表征形式、输入模态、网络结构等实验设置下的有效性与泛化性，并且在四个公开手部数据集上都达到了 state-of-the-art 的效果。

手部 3D 姿态估计问题可以归类于关键点回归任务，目前主流的关键点回归方法可以分为两种：基于回归的方法和基于检测的方法。

（1）基于回归的方法直接学习从深度图到输出关节点坐标的映射，可以端到端训练。这类方法通常在网络输出层用全连接层聚合特征图上每个像素点的信息，用于推导关节点坐标，这样可以捕捉全局信息，但是丢失了特征图上的空间信息；

（2）基于检测的方法通常预测关节点的稠密特征，如高斯热图，其上每个像素点代表这个点是关节点的概率，然后用 argmax 找出峰值对应的索引即为关节点位置。这类方法一般采用全卷积的网络结构，能够保持手的空间结构，但是从高斯热图推导关节点坐标的 argmax 操作是不可导的，因此不能端到端训练，并且当高斯热图的尺寸小于输入尺寸时，存在量化误差。

针对以上两种方法存在的优缺点，研究人员设计了一个自适应加权模块（Adaptive Weighting Regression，简称 AWR），通过一种可导的信息聚合方式，在权重图的指导下，自适应地选择聚合信息的范围，从稠密特征中恢复关节点位置。AWR 模块将稠密特征和关节点位置结合在一起，在不增加额外参数和计算量的前提下，使网络可以端到端训练，同时具备高准确性和高鲁棒性。

A. AWR 自适应加权回归模块

常用的稠密特征有高斯热图（Heatmap）和偏置向量场（Offset field）两种，如下图所示。

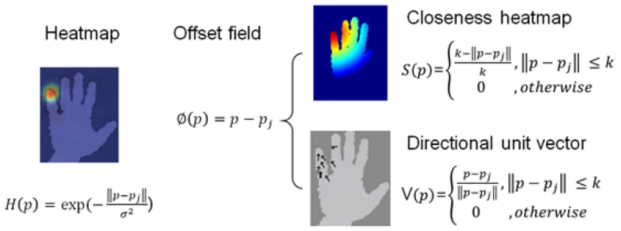


Fig. 5. **稠密特征中的高斯热图和偏置向量场**

AWR 利用权重图对稠密特征的信息进行聚合，这个权重图可以是稠密特征里自带的，例如高斯热图本身和偏置向量场中的 closeness 热图。如下式，AWR 首先对权重图进行归一化，再利用归一化后的权重图对稠密特征图上每个像素点恢复出的关节点坐标进行加权。对于高斯热图来说，每个像素值表示其为关节点的概率，这个概率描述了这个像素值对最终关节点位置推导的

贡献，将高斯热图上每个像素点的贡献加权得到的即为关节点坐标；对于偏置向量场来说，每个像素值的坐标加其相对关节点的偏移向量就是关节点坐标，利用归一化后的权重图加权亦可得到关节点坐标。

$$p_j = \sum_{i=1}^{n} w_j\left(p_{ij}\right) \times j\left(p_{ij}\right)$$

其中 $\left(p_{ij}\right)$ 为关节 j 的密集表示中的 n 个像素之一；$j\left(p_{ij}\right)$ 表示关节 j 中的点 pi 的手关节坐标; $w_j\left(p_{ij}\right)$ 表示点 $\left(p_{ij}\right)$ 的值在权重图中，使用 softmax 函数进行归一化。该操作是可微的，计算简单，能够统一密集表示和手关节坐标。并允许端到端训练，从而缩小训练和推理之间的差距。通过密集表示和联合坐标监督，能够在密集表示中自适应聚合空间信息。

网络训练时先用 dense loss 训练，使权重图收敛到对应的关节点附近，再用关节点 joint loss 进行 finetune，使权重图在遇到关节点遮挡、关节点间存在自相似性等复杂情况时，权重可以发散到相邻的关节点，利用全局信息推导关节点坐标。这样使得网络对各类情况都具备鲁棒性。如下图 7所示食指指尖的权重图，当关节点正常可见时（第一行），AWR 方法得到的权重图和传统基于检测的方法一致，但当关节点被遮挡（第二行）或附近有相似的关节点（第三行）时，权重能发散到相邻的关节点，预测的结果也更为准确。

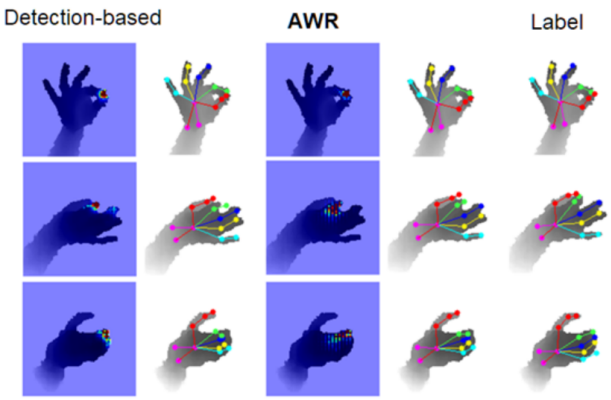


Fig. 6. **食指指尖的权重图**

B. 密集表示类型

AWR 使用了六种最常用的密集表示类型来观察性能。首先，提出了通过堆叠目标关节的 x，y 和 z 坐标来代表一个简单的密集表示。每个关节对应于包含三个通道的姿势图，其中每个通道包含关节相同的 x、y、z 坐标值。由于姿势热图缺少从像素到目标关节的距离的测量，因此会通过无监督学习获取权重图，来聚合局部证据。这种类型的表示称为“P”。

热图是人体和姿势估计中最常用的表示形式，尤其是概率热图，其中每个像素表示特定关节的概率。由于概率热图仅表示关节坐标的 2D 属性，并且 3D 热图的计算成本很高，因此 AWR 引入了深度图或深度偏移图来编码深度信息。深度图和深度偏移图与概率热图具有相同的分辨率，分别表示目标关节的深度值或从当前像素到目标关节的深度偏移。我们将概率热图与深度图相结合称为“H1”，另一个称为“H2”。

接着，使用锚点和手关节之间的 2D 偏移来表示关节的 2D 位置。由于偏移量差异较大，需要将它们分解为 2D 定向单位矢量场和接近热图，反映深度图像中每个像素到目标关节的 2D 方向和接近度。2D 偏移同样缺乏深入的预测，因此，AWR 同时预测每个像素的深度值或从当前像素到目标关节的深度偏移，它们被称为“O1”和“O2”。

但是如果将平面和深度坐标的预测分开，仅使用 2D 距离生成接近热图，那么在一定程度上忽略了深度图像的 3D 属性。因此需要使用 3D 偏移量预测 3D 方向单位矢量场和接近热图，反映 3D 方向和每个像素与目标手关节的接近度。这种方法将三个坐标的预测统一在一起，充分利用深度图像中存在的三维空间信息。在实验部分中将 3D 偏移称为“O3”。

偏移表示 $\phi\left(p_i, p_j\right)$ 的公式如下所示。

$$\phi\left(p_i, p_j\right) = \begin{cases} 1_{\text{Hand}}\left(p_i\right) \times \left(p_i - p_j\right), & \left\|p_i - p_j\right\| \leq k \\ 0, & \text{otherwise} \end{cases}$$

其中 p_i 和 p_j 表示在深度图像和目标手关节中的像素的 2D 或 3D 坐标;k 表示深度图像中的像素到目标手关节的最大距离，$1_{\text{Hand}}\left(p_i\right)$ 是一个指示函数。如果 p_i 属于手部，则 $1_{\text{Hand}}\left(p_i\right)$ 等于 1，否则为 0。

对于 2D 和 3D 偏移，它们进一步分解为二维和三维中的方向单位矢量 $V\left(p_i, p_j\right)$ 和接近热图 $S\left(p_i, p_j\right)$。

$$S\left(p_i, p_j\right) = \begin{cases} 1_{\text{Hand}}\left(p_i\right) \times \frac{k - \left\|p_i - p_j\right\|}{k}, & \left\|p_i - p_j\right\| \leq k \\ 0, & \text{otherwise} \end{cases}$$

$$V\left(p_i, p_j\right) = \begin{cases} 1_{\text{Hand}}\left(p_i\right) \times \frac{p_i - p_j}{\left\|p_i - p_j\right\|}, & \left\|p_i - p_j\right\| \leq k \\ 0, & \text{otherwise} \end{cases}$$

C. AWR 网络架构

AWR 使用简单的 2D CNN 网络，例如 ResNet 或 Hourglas 来提取特征图，然后应用几个反卷积层来提高

提取的特征图的分辨率，并应用几个卷积层来生成密集表示。由于不同类型的密集表示包含不同的组成部分，因此 AWR 使用分离的卷积层来生成它们。

具体来说，对于 O3，使用两个磁头分别输出方向单位向量场和接近热图。而对于 O1 和 O2，我们添加另一个头来生成深度图或深度偏移图。对于 H1 和 H2，分别应用两个卷积层输出概率热图和深度图或深度偏移图。对于 P，两个头分别输出姿势图和权重图，权重图由网络自适应学习，无需监督。

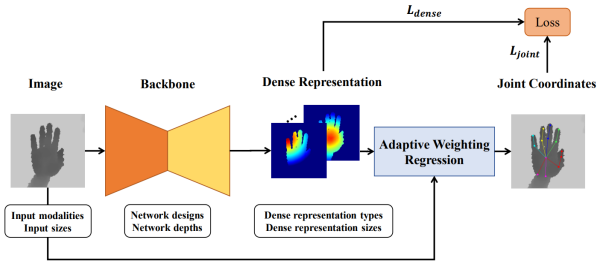


Fig. 7. AWR pipeline

V. 数据集

本次实验的数据集为老师提供的 NYU-part 数据集，该数据集是 NYU Hand Pose 数据集经过处理后的子集。NYU Hand pose 数据集包含 8252 个测试集和 72757 个训练集帧，这些测试集帧包含捕获的 RGBD 数据和地面真实手部姿态信息。对于每一帧，提供 3 个 Kinects 的 RGBD 数据：一个正视图和两个侧视图。训练集仅包含单个人物（Jonathan Tompson）的样本，而测试集包含两个人物（Murphy Stein 和 Jonathan Tompson）的样本，并且每个视图都含有对应的手部姿势的合成重建（渲染）。

本次实验使用的是其经过处理后的子集，该数据集包含捕获的深度图像数据和 ground-truth 手部姿态信息，其中包含了 3000 张训练图片以及 1000 张测试图片，图片为 128*128 的灰度图像。我们可以得到如图 8所示数据集图片样本。

该数据集的标签为描述了手的完全自由度姿态的 42 维向量，其中包含了表示手的 6DOF 位置和方向以及 36 个内关节角度，每个手关节的自由度如图 9所示。对于手部的描述，该数据集使用高质量的线性混合蒙皮（LBS）模型替代了传统的简单半圆形圆柱体模型，虽然 LBS 模型无法准确地模拟肌肉变形和皮肤折叠等效果，但它仍能表现许多球棒模型所无法描述的几何细节，达到更好的效果。

Fig. 8. NYU-part 数据集的图片样本

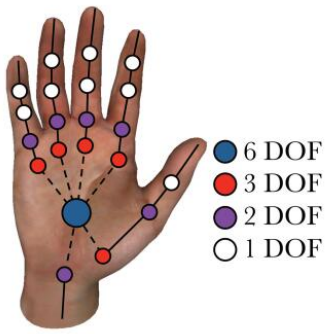


Fig. 9. 具有 42 自由度的线性混合蒙皮 (LBS) 模型

该数据集的产生，是通过基于 RDF 的二分类方法，从背景中分割出手部，并通过直接搜索方法来推导姿态参数。从近似手姿态开始，绘制合成深度图像，并使用标量目标函数与深度图像进行比较。这些深度图像均在基于 OpenGL 的框架中渲染的，其中唯一的渲染输出是与相机原点的距离，且使用的相机也与 PrimeSenseTMIR 传感器具有相同的属性（例如焦距等参数）。

在实际操作过程中，当拟合记录帧序列时，将使用前一帧的姿势来估计手的姿势。而使用序列中第一帧的简单 UI 手动估计姿势，会导致在给定估计的姿态系数的情况下，将由单个标量值表示拟合质量。并且使用带有部分随机化的粒子群优化（PrPSO）直接搜索方法来调整姿态系数值，期望找到最小化该目标函数值的最佳拟合姿态。该算法的概述如图 10所示。

VI. 实验与分析

A. 回归分析

我们开始对该问题的研究之后，首先尝试用老师教授的基础知识进行分析。

输入是 1x128x128 的灰度图，输出是 42 维的向量，我们使用 pytorch 首先构建了一个 resnet50 进行

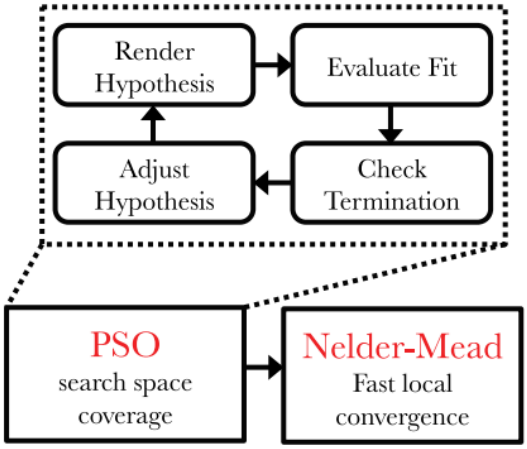


Fig. 10. 用于数据集创建的算法流程

回归分析。优化器使用的是 SGD，损失函数使用的是 MSE，经过 batch_size=32，epochs=10 的训练之后，模型在训练集上的 R^2 为 0.5009891，在测试集上的 R^2 为 0.491298。

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\
SS_{\text{tot}} &= \sum_i (y_i - \bar{y})^2 \\
SS_{\text{reg}} &= \sum_i (f_i - \bar{y})^2 \\
SS_{\text{res}} &= \sum_i (y_i - f_i)^2 = \sum_i e_i^2 \\
R^2 &= 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}
\end{aligned}$$

R^2，决定系数，或称判定系数（英语：Coefficient of determination）在统计学中用于度量应变数的变异中可由自变量解释部分所占的比例，以此来判断回归模型的解释力。不过虽然我们是基于回归的方法在手势估计，但这个评价指标不如 Mean Joint Error 专业。

MPJPE 是“Mean Per Joint Position Error”，即“平均（每）关节位置误差”。MPJPE 常常用于 3D Human Pose Estimation 算法的评价指标，这个指标越小则可认为这个 3D 人体姿态估计算法越好。该值代表的是预测关节点与对应 GT 关节点的 L2 距离的平均值。

$$E_{MPJPE}(f,\mathcal{S}) = \frac{1}{N_{\mathcal{S}}}\sum_{i=1}^{N_{\mathcal{S}}} \|m^{(f)}_{\mathbf{f},\mathcal{S}}(i) - m^{(f)}_{\mathbf{gt},\mathcal{S}}(i)\|_2$$

paperwithcode 榜单上的仓库几乎都使用了 Awesome Hand Pose Estimation 这一仓库中的代码来评估模型性能，故我们使用的也是该仓库中的评估函数，计算得到的 train mean error 为 0.463652，test mean error 为 0.408251。虽然结果相较于 paperwithcode 的榜单 sota 的 6.4 高了特别多，但鉴于原数据集的规模和我们使用的相差甚远，以及我们的标签与原数据集的标程量纲不同，故我们认为这不能代表我们的结果特别好。

在这部分中虽然也可以使用训练 CV 模型常用的 Tricks 来提升模型性能，例如图像增强、更好的模型、学习率和学习率调度器、优化器、正则化手段和知识蒸馏等，但是由于我们对该方面的研究不够全面，不能对问题进行错误分析、消融实验，例如图像增强，我们并不确定增强之后的图片是否适配标签中的自由度坐标，故我们尝试过后放弃了这一改进方向。

Fig. 11. 预测可视化

B. 模型改进

简单的 resnet50 并不能取得让我们满意的结果，故我们参考了 AWR 的基于回归和基于检测的方法，在 resnet 的基础上先提取特征图，然后应用几个反卷积层来提高提取的特征图的分辨率，并应用几个卷积层来生成密集表示。

我们虽然也跑通了 AWR 的官方实现，但实验中使用的是 MMPose 仓库来做相关训练。MMPose 是一款基于 PyTorch 的姿态分析的开源工具箱，是 OpenMMLab 项目的成员之一。主分支代码目前支持 PyTorch 1.5 以上的版本。

MMPose 支持当前学界广泛关注的主流姿态分析任务：主要包括 2D 多人姿态估计、2D 手部姿态估计、

2D 人脸关键点检测、133 关键点的全身人体姿态估计、3D 人体形状恢复、服饰关键点检测、动物关键点检测等。可惜其中还没有包括基于深度图像的 3D 手部姿态估计，故有关 nyu hand dataset 和 AWR 的代码是我们从 github 上下载后修改编写的。

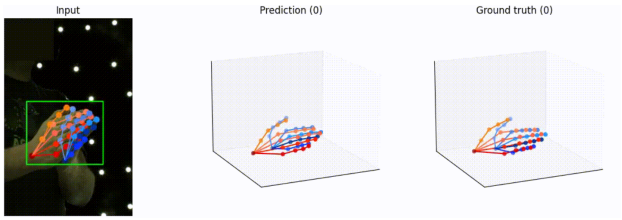


Fig. 12. MMPose 手部姿态估计

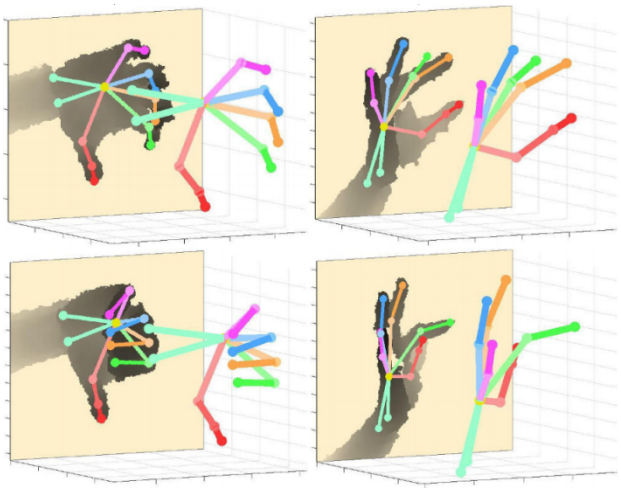


Fig. 13. AWR 预测可视化

在这部分中我们引入了一些额外数据进行训练，该方法在 paperwithcode 榜单上各个仓库中也很常见，除了 uvd 坐标和 xyz 坐标，前者是图像坐标，后者是世界坐标，我们还使用了预先计算的中心。中心坐标数据是通过训练 DeepPrior++ 的手中心估计网络获得的。每条线代表每一帧的三维世界坐标。如果深度图不存在或不包含手，该帧被认为是无效的。

最终计算得到的 Mean Joint Error 为 0.1。

VII. 实验总结

近年来人工智能及其相关领域得到了飞速的发展，人与计算机的交互方式正朝着更自然、更普遍的方向发展。手作为人日常活动的重要组成部分，在人机交互、虚拟现实、机器人等众多应用中是必不可少的组成部分，可广泛应用于娱乐、消费、智能家居、智能驾驶、医疗、工业设计乃至空间应用领域，这使得手势估计受到了人们的广泛关注。手势估计的目的是在三维空间中恢复手部的完整运动姿态，使计算机或者其它设备能够感知人手的空间姿态，从而按照人的指令执行。然而，三维手势姿态估计目前仍然存在很多有待解决的问题，如可获取的手的分辨率低、人手的高自由度、易受环境影响、变化速度快、遮挡和手的相似性等难题都对手势估计的实际应用造成了阻碍。

手势姿态估计在增强现实，虚拟现实以及人机交互与协作等领域存在巨大的应用潜力，一直以来是计算机视觉领域的重点研究方向。传统的机器学习方法大多以 RGB 图像作为输入数据，通过复杂的优化算法拟合手部模型参数。由于手指之间存在着严重的自相似和自咬合问题，缺乏对象深度信息的 RGB 图像往往无法得到准确的估计结果。并且这类传统机器学习方法对于手部快速运动的场景容易丢失跟踪对象，导致其很难适用于实际的应用场景。随着大规模手势姿态数据集的出现和人工智能理论的发展，基于深度图像的卷积神经网络方法逐渐成了手势姿态估计的主流方法。当下的众多方法往往通过构建复杂的三维神经网络，致力于提升单一的估计精度，而忽略了模型复杂度过大带来的推理效率低下的问题。因此，本次作业中我们以回归分析为主，再加上基于检测的方法辅助，使用 AWR 在保证较高估计精度的同时实现更高的实时推理速度。

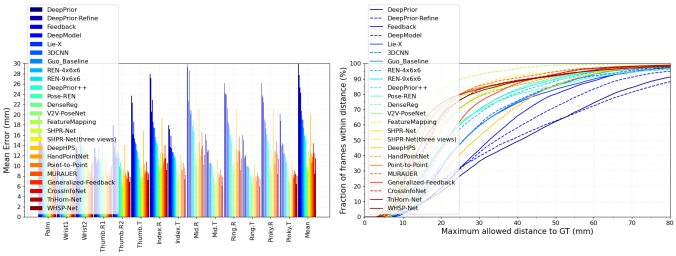


Fig. 14. NYU hand dataset：Mean error for each joint

如图 14所示，Hand Pose Estimation 这一任务非常热门，有非常多的科研工作者奉献其中。本次实验中我们复现了 paperwithcode 上 nyu hand dataset 几乎所有的开源仓库，虽然由于跟我们的数据子集差异较大起初遇到了很多困难，但是在解决问题的过程中我们也受益匪浅。相信在未来基于深度图像的手部姿态估计的算法会继续取得令人瞩目的进展，在公开的基准数据集上的最佳纪录将不断被打破。