红外图像下的手部姿态估计

郑晓豪
SZ2316132
计算机科学与技术学院

日期：2024 年 5 月 25 日

摘　要

手部姿态估计在计算机视觉领域中具有重要性，广泛应用于人机交互、手语翻译、增强现实、机器人控制以及游戏娱乐等领域。通过精确识别手部姿态，可以实现自然直观的人机交互，提高沟通效率，增强用户体验，并提供创新的解决方案。红外图像在手部姿态估计中具有独特优势。不受光照变化影响，能够在各种光照条件下稳定工作，同时具有隐私保护、高对比度、抗干扰能力强等特性。红外图像还能提供温度信息，增加估计的准确性和鲁棒性。在本文中，提出基于 RTMPose 改进的关键点检测模型 IRHPose 和基于温度的手部判断模型 TB-HJM。IRHPose 利用多分支图卷积根据先验知识将手的结构融合进模型。TB-HJM 用于选择与温度分布最接近的特征图作为最终输出，解决红外图像中存在视觉信息不足的问题。和基准模型相比，本文提出的模型在红外手势数据集上取得最好的性能。

关键词: 深度学习，手部姿态估计，红外图像，关键点检测

1　引言

手部姿态估计是计算机视觉和模式识别领域的重要研究方向之一，涉及通过分析图像或视频数据来确定手部的姿态和动作。其应用领域非常广泛，包括人机交互、虚拟现实（VR）、增强现实（AR）、手语翻译、医疗康复、机器人控制和游戏娱乐等。在人机交互领域，通过精确的手部姿态识别，可以实现自然直观的用户界面控制；在医疗康复中，手部姿态估计能够帮助患者进行康复训练，监测其康复进度；在手语翻译中，精确的手部姿态识别对于促进听障和语障人士的沟通具有重要意义。

红外图像在手部姿态估计中具有许多独特的优势。首先，红外图像不受光照变化的影响，能够在各种光照条件下稳定工作，这对于光线变化剧烈的应用场景尤为重要。其次，红外图像具有高对比度，便于手部区域的分割和特征提取。此外，红外图像在保护个人隐私方面具有优势，因为它不会捕捉详细的可见光特征。红外图像还具有较强的抗干扰能力，不受可见光范围内的干扰源影响。最后，红外图像可以提供温度信息，利用手部的热图特征进一步提高姿态估计的准确性和鲁棒性。这些优势使得红外图像在手部姿态估计中展现出巨大的潜力。

本文旨在研究利用红外图像进行手部姿态估计的方法，以克服现有方法在光照变化、视觉信息少和抗干扰能力方面的不足。具体来说，本文的主要贡献包括：

1). 创建一个用于红外手势图像数据集，命名为 InfraredHands 数据集。本文中所有的实验均在该数据集上进行。

2). 介绍了一种名为 IRHPose 的手部关键点检测模型，该模型在红外图像中表现出高精度，通过多分支图卷积根据先验知识将手的骨骼结构融合进模型。

3). 构建了一个基于温度的手部判断模型（TB-HJM），该模型包含了手部关键点潜在温度分布的信息。结合手的温度信息来解决红外图像中存在视觉信息不足的问题。

2 相关工作

2.1 人体姿态估计

由于手部姿态估计任务与一般人体姿态估计任务具有相似性，并且专门针对手部姿态估计的研究较为有限，讨论二维人体姿态估计的相关工作。人体姿态估计是人体动作识别中的上游任务。目前，大多数二维人体姿态估计任务可以根据输出表示的不同分为两类：基于回归的姿态估计和基于热图的姿态估计。接下来，介绍这两类 HPE 方法的相关研究。

1). 基于回归的人体姿态估计

早期的回归方法主要是通过直接回归图像到人体关键点的位置。Toshev[1] 和 Szegedy 提出了 DeepPose 模型，利用深度神经网络直接回归人体关键点的坐标。这种方法的优点在于简单直接，但在面对姿态变化较大和背景复杂的情况下，效果有限。为了提高回归的精度，一些研究提出了级联回归的方法。Carreira[2] 等人提出了 Iterative Error Feedback (IEF) 方法，通过级联多层回归器逐步优化关键点位置。Sun[3] 等人则提出了 Hourglass 网络，通过多层次的特征提取和逐步细化，显著提升了姿态估计的准确性。

2). 基于热图的人体姿态估计

热图回归方法通过预测每个关键点在图像中的概率分布（即热图），然后从热图中提取关键点位置。Tompson[4] 等人首次提出了这种方法，通过卷积神经网络预测每个关键点的热图，并使用空间软化策略来提高精度。Newell[5] 等人进一步提出了 Stacked Hourglass Networks，通过多尺度特征融合和反复的上下文信息聚合，实现了更高精度的姿态估计。为了更好地捕捉人体的多尺度特征，一些研究引入了多尺度特征融合技术。Chen[6] 等人提出了 Cascaded Pyramid Network (CPN)，通过级联金字塔结构逐步细化关键点位置预测。Xiao[7] 等人则提出了 Simple Baseline 方法，通过简单但有效的上采样层实现了高分辨率的热图预测。自注意力机制在人体姿态估计中的应用也取得了显著进展。Sun[3] 等人提出了 HRNet（High-Resolution Network），通过保持高分辨率特征图的计算，实现了更精细的热图预测。Li[8] 等人则在 HRNet 基础上引入了自注意力机制，进一步提升了姿态估计的性能。

2.2 红外图像在计算机视觉中的应用

红外图像在计算机视觉中的应用广泛，主要利用其在不同光照条件下的稳定性和独特的热图特征。在行人检测、夜视监控、医学影像分析等领域，红外图像展现出显著的优势。Hossain[9] 等人研究了红外图像在夜间行人检测中的应用，通过热图信息实现高精度的行人识别。Xu[10] 等人在医学影像分析中利用红外图像监测人体的温度分布，辅助疾病诊断和治疗。

2.3 红外手部姿态估计方法

在手部姿态估计领域，红外图像的应用也得到了越来越多的关注。主要包括以下三种：红外深度融合、基于红外热图特征和红外图像增强。红外深度融合方法：为了充分利用红外图像和深度图像的互补特性，一些研究将二者进行融合。Zimmermann[11] 等人提出了一种融合红外和深度图像的手部姿态估计方法，通过双通道 CNN 同时处理两种图像，提高了姿态估计的精度。基于红外热图特征的方法：一些研究直接利用红外图像的热图特征进行手部姿态估计。例如，Sridhar[12] 等人提出了一种基于红外热图的手部姿态估计方法，通过热图特征提取手部关键点位置，结合深度学习网络进行姿态预测。红外图像增强技术：为了提高红外图像的质量和特征提取的有效性，一些研究引入了图像增强技术。Wang[13] 等人通过红外图像增强和去噪技术，提升了红外图像的对比度和细节保留，为手部姿态估计提供了更优质的输入数据。

3 方法

3.1 红外图像手部关键点检测模型

RTMPose 是一款通用的关键点检测模型，而本文的研究场景是基于红外图像的手部关键点检测，因此 RTMPose 模型在该场景下的局限性在于没有考虑到手部的结构先验。本文在 RTMPose 的基础上通过添加多通道卷积结构将手部结构先验隐式的融入到网络中。同时，本文通过在 RTMPose 的头结构中增加图卷积神经网络模块（Graph Convolutional Network Module, 简称 GCN 模块），利用手部关节点的相连关系进行建图，将手部结构先验显示的融入到网络中。通过这两种方式能够很好地将手部结构先验融入到模型当中，经过实验验证，融入了多通道卷积结构以及 GCN 模块后的 RTMPose 在精度上有较大提升。如图 1 所示，根据手天然的物理结构，将 21 个关键点分为五组。关键点之间的相关性决定分组，分到同一组的关键点相关性高，不同组中的关键点相关性低。如图 2 所示，数据通过 BackBone 后，输入到多分支卷积模块中进行分组预测。不同组之间的采用不同的分支进行预测，通过这种方式能够做到让相关性高的关键点共享大量特征，相关性低的关键点之间共享更少的特征。

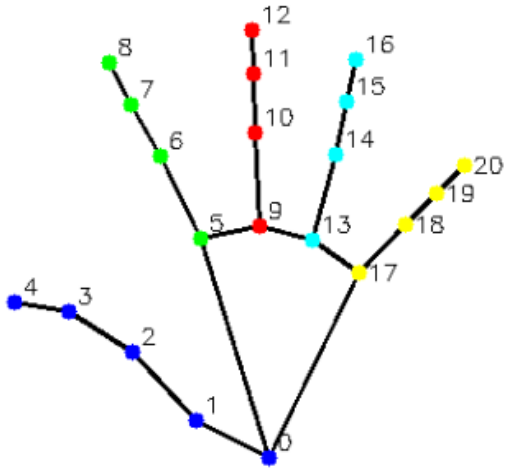

图 1: 手部关键点结构分组图

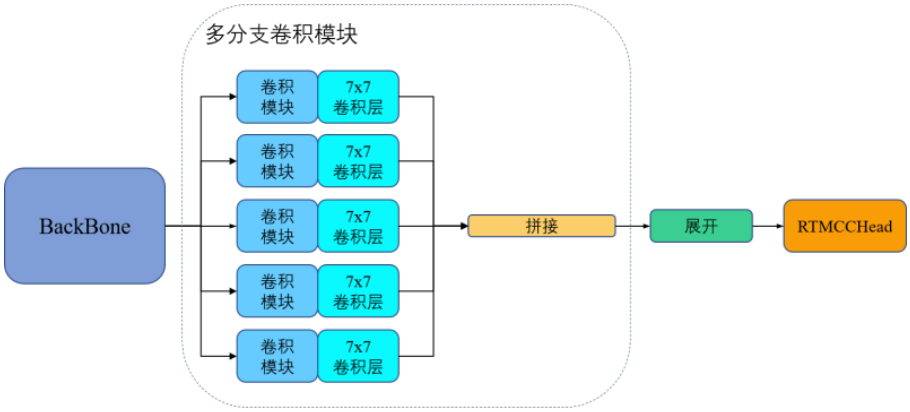

图 2: 多分支图卷积结构图

红外图像的手部关键点检测模型的网络结构如图 3 所示。首先还是经过骨干网络提取特征，再经过多分支卷积模块为每个关键点生成特征矩阵，经过展开操作后生成一维的特征向量，随后经过一个全连接层进一步提取特征。图卷积神经网络每一层的节点都融合了与之相邻的所有节点的信息，因此在全连接层后拼接上三个残差语义图卷积模块作为 Head 部分的主干网络。包含多个图卷积层可以保证图中所有节点信息都能通过事先定义的图结构充分融合。在每个残差语义图卷积模块后都拼接上一个 RTMCCHead 用于输出中间结果。

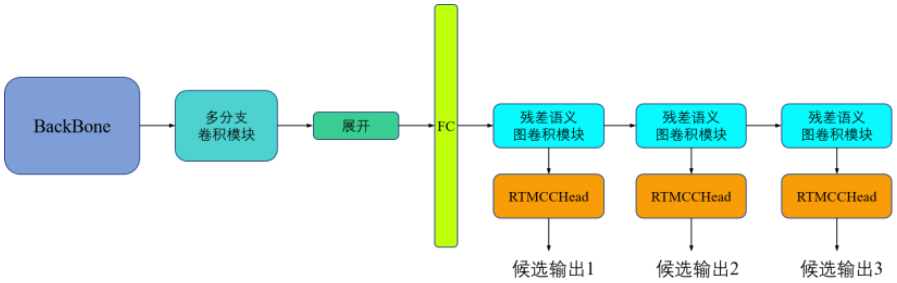

图 3: IRHPose 模型结构图

3.2 基于温度的手部判断模型

红外图像是灰度图像且灰度值与温度具有强相关。因此手部在红外图像下会呈现某种特定的灰度分布。但是检测手部的灰度分布特征相对来说是一个比较困难的任务。因此选择 21 个关键点对应位置的灰度数据作为整个手部灰度分布的表征，通过一个简单的深度学习模型可以判断输入的 21 个灰度值是否符合潜在的红外图像下手部灰度分布。借助这个判别模型可以轻易地从 IRHPose 输出的三个手部候选中选择一个最符合手部灰度分布特征的手部候选作为最终的输出。如图 4 所示为 TB-HJM 的网络结构图。

本文设计了基于 TB-HJM 的动态权重损失函数，该损失函数通过对不符合灰度分布的手部预测增加惩罚项使得模型梯度向更优的方向下降。本文提出来的 IRHPose 一共会输出三个手部候选，因此需要对每个手部候选分别计算 KL 损失。

公式中的代表 Q_i 第 i 个手部候选的 x 轴与 y 轴分类标签的概率分布。是 TB-HJM 的输出，P_i 代表第个手部候选是一个手的概率。每个手部候选的损失都由原始的 KL 散度以及基于 TB-HJM

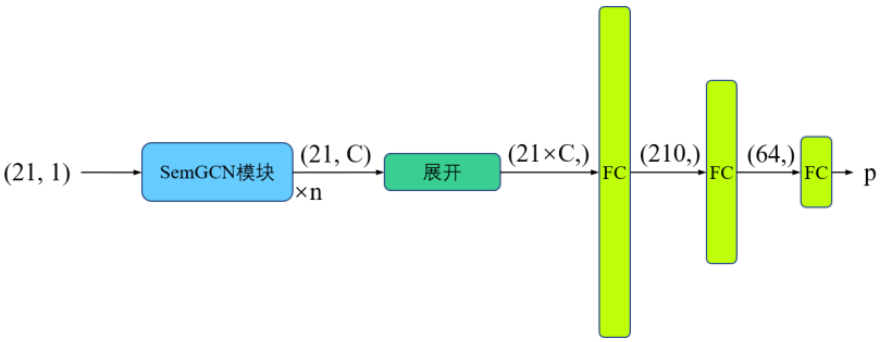

图 4: TB-HJM 结构图

输出加权的 KL 散度两项组成。第二项用于对不符合灰度分布的预测进行惩罚。

$$\text{Loss} = \frac{1}{3}\sum_{i=1}^{3} \text{Loss}_{KL}(Q_i) + (1 - P_i) \cdot \text{Loss}_{KL}(Q_i) \tag{1}$$

4 实验结果

4.1 数据集构建

采集设备为智能手机热像仪 Infray。数据集共包含 6756 张带标注红外手势图像，由九名实验人员共同采集数据，包含十六种以上常见的军用手势。不同成员采集的手势类别并不固定，且不同成员采集的样本数量也不固定。实验人员的手部特征均不相同，基本能够涵盖绝大多数特征的手部。数据集中的手势，基本能够涵盖绝大多数特种兵手语交流场景。所有手势的带标注图像都在 300 张以上，能够保证每种手势都有足够的带标注数据。如图 5 所示，为部分红外手势图像。

图 5: 部分红外手势图

4.2 实验结果与分析

实验的软件环境为 Ubuntu 20.04 操作系统，Python 3.8、PyTorch 2.1、CUDA 11.8。实验结果如表 1 所示,PCK 指的是正确预测的关键点比例，PCK@0.02 代表归一化距离为 0.02 时的 PCK 值，即在归一化距离为手部尺寸的 2% 时，正确关键点的百分比。EPE 指的是端点误差，即预

测关键点位置与真实位置求欧式距离再取平均。在同等量级的计算量与参数量下，本文提出的IRHPose 精度明显高于其它模型，如 $IRHPose_s$ 在计算量仅为 $RTMPose_l$ 五分之一，参数量不到 $RTMPose_l$ 一半的情况下取得了与 $RTMPose_l$ 近似的精度（AUC 提高了 0.11%，PCK@0.02 降低了 0.24%，EPE 降低了 0.03）。

表 1: IRHPose 与现有模型的对比实验

model	flops(G)	params(M)	AUC(%)	PCK@0.02(%)	EPE
HRNet	10.248	28.536	83.02	81.19	4.37
SBL_r18	3.876	15.378	82.35	78.56	4.6
SBL_r50	7.271	34.001	82.98	80.15	4.39
RTMPose_l	5.559	27.898	85.03	82.3	3.79
RTMPose_m	2.581	13.775	84.85	81.68	3.82
RTMPose_s	0.916	5.61	84.33	80.86	4.06
IRHPose_l	5.92	34.876	85.72	83.23	3.56
IRHPose_m	2.864	19.542	85.58	82.68	3.6
IRHPose_s	1.122	10.168	85.14	82.06	3.82

通过分析可知，IRHPose 能达到更高精度原因在于以下三个方面：1）通过多头多分支头部设计与 GCN 模块将手部结构先验融入到模型当中；2）通过 TB-HJM 监督模型训练同时在多个手部候选中选择一个更优结果进行输出；3）通过红外图像信息增强预处理对图像信息进行增强。如图如图 6 所示为 IRPose 的输出结果图。

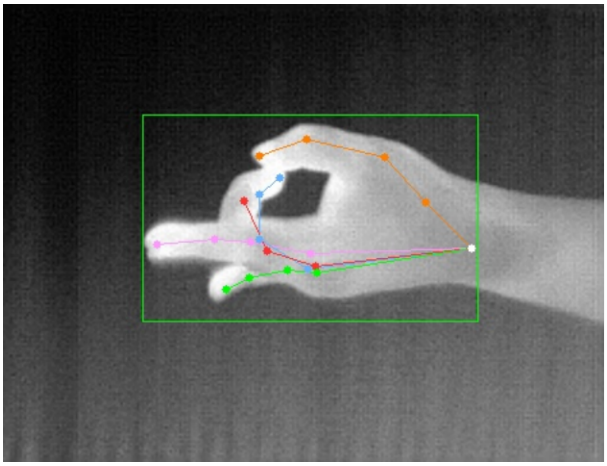

图 6: IRPose 输出结果图

5 结论

本文创建一个用于红外手势图像数据集，命名为 InfraredHands 数据集，丰富了手势数据集的种类。还提出一种通过多分支图卷积根据先验知识将手的骨骼结构融合进模型的 IRHPose 的手部关键点检测模型。同时通过基于温度的手部判断模型（TB-HJM），结合手的温度信息来解决红外图像中存在视觉信息不足的问题。本文提出的模型在红外图像手势数据集上取的比其他

方法更好的效果。未来的研究方向是将红外图像中的手部姿态估计任务与红外图像下的手势识别任务相结合，实现端到端的红外动态手势识别。

参考文献

[1] Alexander Toshev and Christian Szegedy. “DeepPose: Human Pose Estimation via Deep Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pp. 1653–1660.

[2] Joao Carreira et al. “Human Pose Estimation with Iterative Error Feedback”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 4733–4742.

[3] Ke Sun et al. “Deep High-Resolution Representation Learning for Human Pose Estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 5693–5703.

[4] Jonathan Tompson et al. “Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2014, pp. 1799–1807.

[5] Alejandro Newell, Kaiyu Yang, and Jia Deng. “Stacked Hourglass Networks for Human Pose Estimation”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2016, pp. 483–499.

[6] Yilun Chen et al. “Cascaded Pyramid Network for Multi-Person Pose Estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 7103–7112.

[7] Bin Xiao, Haiping Wu, and Yichen Wei. “Simple Baselines for Human Pose Estimation and Tracking”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 466–481.

[8] Wenzhe Li, Zongxin Wang, and Shengping Yang. “Multi-Scale Attention Network for Human Pose Estimation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 6460–6469.

[9] Mohammad M Hossain and Govinda Chetty. “Night Vision for Pedestrian Detection with FIR: A Review”. In: *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2015, pp. 1–7.

[10] Xianzhong Xu, Hong Guo, and Jiali Han. “Infrared Image Analysis for Medical Applications: A Review”. In: *IEEE Access* 7 (2019), pp. 178518–178531.

[11] Christian Zimmermann and Thomas Brox. “Learning to Estimate 3D Hand Pose from Single RGB Images”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 4903–4911.

[12] Srinath Sridhar, Franziska Mueller, and Antti Oulasvirta. “Fast and Robust Hand Tracking Using Detection-Guided Optimization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 3213–3221.

[13] Robert Y Wang and Jovan Popovic. “Real-Time Hand-Tracking with a Color Glove”. In: *ACM Transactions on Graphics (TOG)* 28.3 (2009), pp. 1–8.