2024 年第六届全球校园人工智能算法精英大赛

基于无人机的人体行为识别任务的解决方案

林宜鹏[1)] 左玲玲[1)] 杨杨[1)]

[1)] (南京理工大学计算机科学与工程学院,江苏省南京市，210094)

摘要 本报告详细介绍了“基于无人机的人体行为识别”竞赛中的算法解决方案，提出了一种结合图卷积网络和多模态融合的创新方法，旨在提升人体骨架行为识别的准确性。首先，我们分析了数据集中的关联性动作问题，并提出了一种基于双人 GCN 建图的方法，通过扩展邻接矩阵来捕捉两人之间的交互关系。其次，为避免冗余信息的干扰，我们引入了 DeGCN 中的可变形关节采样方法，优化了关节建图过程，筛选出与动作最相关的关节信息。此外，我们采用了 SkateFormer 模型的分区自注意力机制，分别建模骨架和时序数据之间的关系，能够有效聚焦于与动作识别最相关的关节和时间帧。最后，采用多模态数据融合策略，并集成 TeGCN、CTRGCN、DeGCN 和 SkateFormer 等多个模型的预测结果，从而显著提高了识别准确性。实验结果表明，所提出的方法在验证和测试数据集上分别达到了 52.8%和 49.71%的准确率，表现出较强的识别能力和鲁棒性。

【关键词】人体行为识别；关联动作建图；可变形关节采样；分区自注意力机制；多模态融合

1 介绍

随着无人机技术的快速发展，其在公共安全、智能监控、灾害救援等多个领域的应用前景日益广阔。在安防领域，利用无人机进行人体行为识别，不仅能够实时监测特定区域内的人体行为，还能有效提升社会管理效率和公共安全。然而，面对复杂多变的实际场景，无人机在人体行为识别任务中仍面临诸多技术挑战。

本次竞赛的目标是利用无人机搭载的传感器和摄像头对人体行为进行准确识别，涵盖但不限于站立、行走、奔跑、跌倒等行为。参赛者需基于一个经公开验证的大型无人机人体行为识别数据集展开挑战。该数据集的 joint 模态为唯一的原始数据来源，是通过无人机摄像头获取的关键关节信息，基于此数据，可以生成补充模态（如 bone 模态、motion 模态等），以提高识别精度和模型的全面性。数据格式为（N，C，T，V，M），其中 N 为样本

数量，C 为通道数，T 为帧数，V 为节点数，M 为坐标；在多人场景中，多个非零坐标对应多个个体。标签格式为（N,），每个样本与一个行为标签对应，任务的难点在于如何有效处理骨骼数据并提升识别效果。

针对这一任务，我们提出了一种创新的算法解决方案，旨在通过多模态融合与图卷积网络相结合，提升人体行为识别的准确性。我们的主要创新点包括：

（1）基于关联动作的建图方法：针对数据集中的关联性动作（如两个个体的互动行为），我们提出了一种改进的 GCN 建图方式，将单一个体的图拓展为两人的交互图，捕捉两人之间的互动关系。传统方法通常将两个人视作独立个体，容易忽略动作间的关联性。通过扩大邻接矩阵并设计有效的交互图，我们能够更好地学习关联动作特征。

（2）可变形关节采样：为了解决关联动作建图后冗余信息过多的问题，我们结合了 DeGCN 模型的可变形关节采样方法。该方法通过确定动作相关的关键关节点，并优化选择与动作最相关的关节信息，有效避免了冗余信息对模型训练的干扰，从而提高了识别精度。

（3）骨架-时序自注意力机制：通过划分不同类型的骨架-时序关系（如邻近与远距离的关节关系、邻近与远距离的帧关系），在每个划分内进行注意力计算，这样可以在不同的时空维度上自适应地选择性关注关节和帧的关键特征。

（4）多模态融合与多模型集成：为了进一步提升识别效果，我们在模型中引入了多模态融合策略。通过对 joint、bone、motion 等多种数据模态进行加权融合，并结合多个图卷积网络模型（TeGCN、CTRGCN、DeGCN、SkateFormer）的预测结果(图 1)，显著提高了识别的准确率。

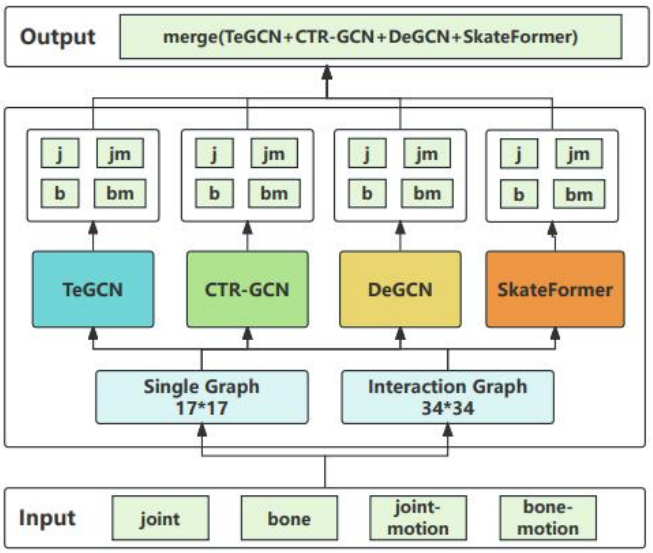


图 1 总体策略

- TeGCN（Temporal Graph Convolutional Network）：该模型专注于时间序列数据的建模，在图卷积过程中引入时间维度，能够有效捕捉人体动作的动态演变特征。
- CTR-GCN（Channel-wise Topology Refinement Graph Convolutional Network）：通过动态学习不同的图拓扑结构并优化通道特定的关节特征聚合，结合共享拓扑结构和通道特定调整，有效减少了建模复杂性。
- DeGCN（Deformable Graph Convolutional Network）：通过自适应地捕捉人体骨骼序列中最具信息性的关节，结合空间和时间图的可变形采样，减少了冗余信息。
- SkateFormer (Skeletal-Temporal Transformer): 该模型是一种结合骨架-时序自注意力机制的变压器模型，通过划分不同类型的骨架-时序关系，能够高效地捕捉关节与帧之间的关键特征。

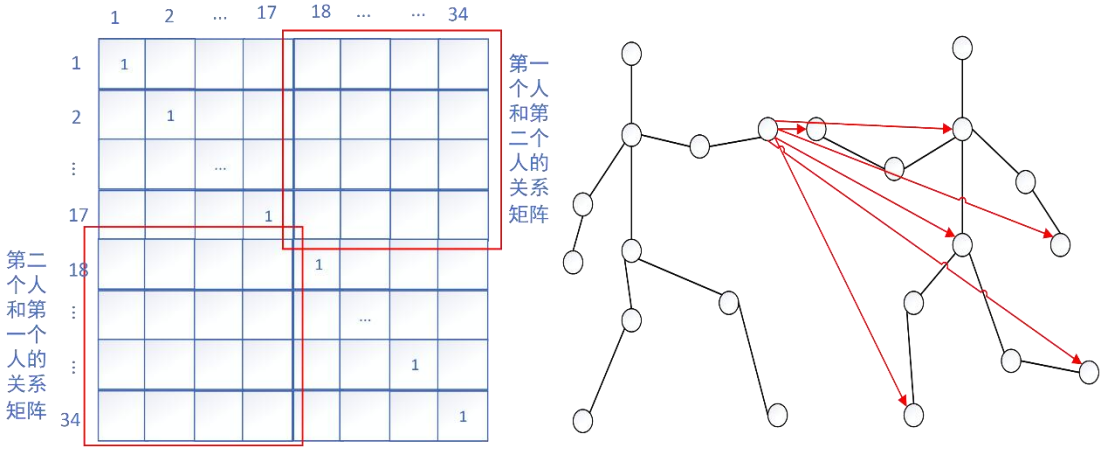
实验结果表明，所提出的方案在复杂场景和多人交互行为识别中具有优越性，在验证集上准确率达到了 52.8%，在测试集上准确率达到了 49.71%。这一结果不仅证明了我们方法在实际应用中的有效性，也为图卷积网络在人体行为识别中的进一步应用提供了新的思路和实践经验。

2 方法

在这节中主要说明我们的解决方案。

2.1 关联动作建图

我们在前期工作中，对比赛的数据集进行了大量的分析工作。在对类别的探索中，我们发现，UAV-Human 的数据集类别中存在着不少关联性动作(比例:32/155)，即需要两个人才能表示的动作，如 A096: exchange something with someone、A082: rob something from someone 等。由于 UAV-Human 大部分单条数据都是包含两个人的关节位置信息，类别标签是对着两个人的动作的唯一标记，目前大部分模型，如 TeGCN、CTRGCN 等在处理骨骼数据时，通常会将在将数据输入模型前，将两个人看作是两个独立的个体，输出在 logit 层时通过对两个人的结果取均值，实现唯一标签的输出。这样的模型流程直接将两个人互相孤立了，模型很难在一定程度上学习到关联性动作的特征，同时这些特殊的关联性动作容易影响模型对独立动作的学习。在此基础上，我们设计了一种新的 GCN 建图方式，将图的结点数量从原来的 17*17 变成 34*34。即从一个人的图拓展到两个人的图。



(a) (b)

图一 两人 GCN 关系建图

如图一(a)所示，我们将 GCN 的邻接矩阵扩大到原来的两倍，左上角和右下角和原来单独建图的方式一致，左下角和和右上角是对于两人之间可能的交互点建图的。由于太多的关联可能导致模型难以学习到真正的动作关系，因此我们首先选取了较少的点关联。图一(b)中描述的是图一(a)中右上角建图的部分关系，展示了第一个人的手腕关节点和第二个人的主要关节点(手脚和驱干)的相连。另外，还需要将第一个人的另外的手腕关节和两个脚腕关节点和第二个人的主要关节点相连。

2.2 可变形关节采样去除冗余

在建图后进行初步实验，我们发现，这种建图方式在 Tegcn 后的准确率提升不是很大，但是在和原来的建图方式进行后融合会很大提升准确率。表 1 中的四个模态分别是 joint、bone、joint_motion 和 bone_motion，v 表示建图的骨骼关节点数。

表 1 关联动作在 Tegcn 上的两种建图方式的准确率

方法	**正确**	**错误**	**Acc(所有类别上)**
4_modality(v=17)	191	179	47.3
4_modality(v=34)	**202**	**168**	47.1
Merge	199	171	**48.5**

如表 1 所示，关联动作建图的方式会在关联动作类上的正确和错误数量都有优化，但是可能会一定程度损害其他单人动作的准确率。我们初步认为这是由于建图后，骨骼关节之间的关联太过密集，存在过多冗余信息，模型难以学习到正确的特征表示。

为了解决这个难题，我们在寻找对应解决策略的过程中找到了 DeGCN。DeGCN 提出人类行为具有高度阶级内差异，即不同的骨架外形表示的是一个动作类别，例如站着看书

和坐着看书都是看书。因此 DeGCN 提出了一种可变形的采样框架，如图 2 所示。

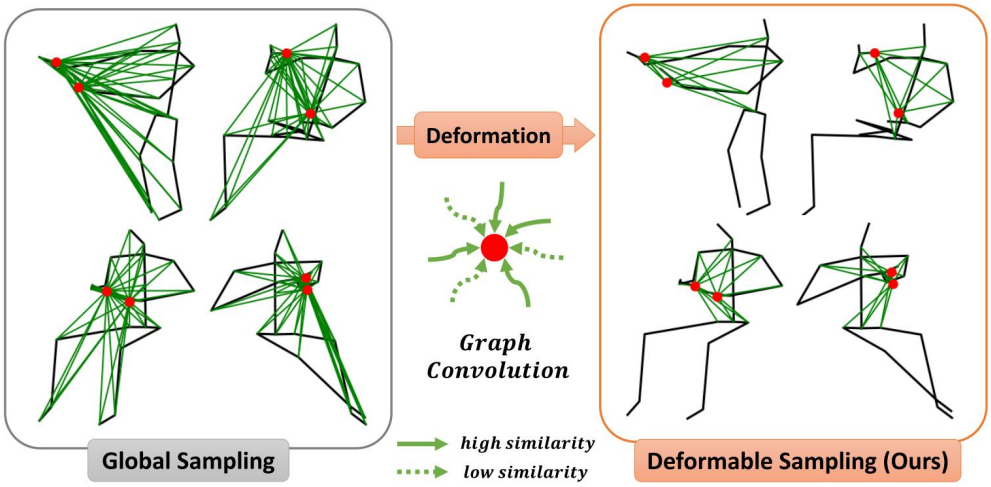


图 2 DeGCN 的可变形关节

图 2 左边是传统方法选取的点，右边是 DeGCN 的可变形关节，可以看到针对当前动作，传统的方式存在很多多余的信息，容易干扰模型的学习。这种方式通过卷积后先确定骨架的中心点，再计算两两关节点的相似度，优先选取相似度分数高的点，并通过损失函数梯度下降不断优化中心点的选取和有用的关节点的筛选，滤除过多的冗余信息。

我们认为这种基于中心点去筛选动作相关的方式能够对于关联动作建图有很大的帮助，因为我们为了描述两个人之间的动作关系，再本就密集的邻接矩阵上添加了新的信息，DeGCN 可以很好的平衡关联动作和独立动作，从而做到模型在独立动作学习中不会将关联动作的连接信息选中，也能以较好的方式学习关联动作。因此我们将关联动作建图的方式融入到 DeGCN 中，最终实验表示模型在 UAV-Human 上面的准确率得到了很大提升。

2.3 骨架-时序自注意力机制

为了进一步提升模型的表达能力，我们采纳了一种创新的骨架-时序 Transformer (SkateFormer) 方法，该方法采用分区自注意力机制，专门针对骨架数据和时序数据之间的关系进行建模。具体来说，SkateFormer 将关节点和帧信息根据不同的时空关系进行分区，并在每个分区内执行自注意力机制。通过这种方式，模型能够更高效地聚焦于与动作识别最相关的关节和时间帧，进一步优化了模型的计算效率与准确性。

3 实验

3.1 实验设置

我们在 CTRGCN、TeGCN 和 DeGCN 分别进行了八个模态的训练，包括 GCN_v=17 和 GCN_v=34 的各四个模态(joint、bone、joint_motion、bone_motion)。最后再把每个模型的

结果按照一定比例融合起来。实验表格中出现的分数，均为测试集的分数，验证集需要提交，我们的验证次数不足以支撑做太多实验。

3.2 对比实验

表 2 UAV-Huamn 数据集在 TeGCN 上的表现

序号	模态	GCN_size	分数
1	joint	17	43.9
2	bone	17	42.5
3	joint-motion	17	34.8
4	bone_motion	17	35.6
5	joint	34	43.9
6	bone	34	41.8
7	joint-motion	34	33.8
8	bone_motion	34	35.7
9	Merge(1-4)	17	47.3
10	Merge(1-5)	/	48.15
11	Merge(1-8)	/	48.5

如表 2 所示，加入了关联动作建图后，仅仅只合并了 joint 的模态，就让模型的准确率提升了很多。为了方便，我们后续的关联动作建图只在 joint 模块上实现。

表 3 UAV-Human 数据集在 CTRGCN 上的表现

序号	模态	GCN_size	分数
1	joint	17	44.05
2	bone	17	45.15
3	joint-motion	17	38.2
4	bone_motion	17	36
5	joint	34	45.55
6	Merge(1-4)	17	48.2
7	Merge(1-5)	/	49.15

表 4 UAV-Huamn 数据集在 DeGCN 上的表现

序号	模态	GCN_size	分数
1	joint	17	47.15
2	bone	17	45.15
3	joint-motion	17	41
4	bone_motion	17	36
5	joint	34	48.4

6	Merge(1-4)	17	48.2
7	Merge(1-5)	/	50.45

可以明显的看到 DeGCN 的准确率远大于 CTRGCN 和 TeGCN，实验结果很好的解释了可变性关节点选择策略的有效性。同时，关联动作建图在这种可变性关节选择策略上的性能达到了前所未有的优秀，仅仅一个 joint 模态的效果就比肩 TeGCN 五个模态融合的效果，证明了关联动作建图的有效性，同时也证明了 DeGCN 设计的初衷—太过冗余的关节信息会阻碍模型的学习。

另外在 SkateFormer 上运行 UAV-Human 数据集时，我们对关节点进行了扩充，从 17 个增加到 25 个。这些调整在不同模态下的表现如下表 5 所示：

表 5 UAV-Huamn 数据集在 SkateFormer 上的表现

序号	**模态**	**GCN_size**	**分数**
1	joint	25	46
2	bone	25	45.3

最后，我们融合了四个模型的分数，如表 6 所示。其中单独模态融合的方式会更优秀，可能是由于太多模态和模型的参与，存在了部分冗余参数。

表 6 模型融合结果

序号	**方法**	**融合方式**	**分数**
1	DeGCN+CTRGCN+TeGCN	各个模态参与	51.2
2	单独模态融合(3*5+2)	[3.6, 5.2, 0.4, 0, 4.8] + [4, 5.6, 0.4, 0, 3.8] + [3.8, 0., 0.2, 0, 0.] + [0.4, 0.2]	52.95

4 总结和展望

在本篇文章中，提出了一种新的基于动作交互的关联动作建图模式，同时在引入了可变性的骨架关节选择模型 DeGCN，在单模态上就达到了很好的性能，极大的提升了模型在 UAV-Human 上的表现。

同时我们在比赛的过程中，基于对 Tranformer 在学习的过程中，会逐渐忽略原始固定的邻接矩阵在模型中的作用的思想，尝试过使用可学习的邻接矩阵，可能是由于时间不足，模型最终性能并没有达到预期的效果。