计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009698] [CSTR: 32024.14.csa.009698]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于全局-个体特征融合的群体行为识别①

程　勇[1], 程　遥[1], 王　军[1], 杨　玲[1], 许小龙[1], 高园元[1], 张开华[2]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](南京信息工程大学 计算机学院, 南京 210044)
通信作者: 程　勇, E-mail: yongcheng@nuist.edu.cn

摘　要: 群体行为识别是计算机视觉领域中备受关注的研究方向之一, 旨在通过多个个体动作与互动关系确定整体的行为. 然而, 由于确定个体互动关系、联系紧密程度以及活动关键人物三者的困难, 现有方法常关注于人物的个体特征, 忽略了与活动场景上下文的相互联系. 针对该问题, 提出一个基于全局-个体特征融合的群体行为识别推理模型 GIFFNet (global-individual feature fusion network). 通过构建全局-个体特征融合 (GIFF) 模块, GIFFNet 在聚焦关键信息的基础上, 有效整合了场景上下文与个体人物特征, 获取了更具表征能力的融合特征, 以弥补预测群体行为时场景信息缺失的问题. 随后, GIFFNet 利用融合特征计算场景中人物之间的交互关系图, 并使用图卷积网络 (GCN) 进行训练和群体行为类别预测. 此外, 为解决数据集样本失衡的问题, GIFFNet 采用动态分配权重的策略优化损失函数. 实验结果表明, GIFFNet 在 Volleyball、Collective Activity 数据集上的多类分类准确度分别为 93.8%、96.1%, 类平均精确度分别为 93.9%、95.8%, 优于其他现有的深度学习方法. GIFFNet 通过特征融合为行为分类提供了表征能力更加强大的特征, 有效地提升了行为识别的精确度.

关键词: 群体行为识别; 场景上下文; 特征融合; 注意力机制; 动态损失函数

引用格式: 程勇,程遥,王军,杨玲,许小龙,高园元,张开华.基于全局-个体特征融合的群体行为识别.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9698.html

Group Activity Recognition Based on Global-individual Feature Fusion

CHENG Yong[1], CHENG Yao[1], WANG Jun[1], YANG Ling[1], XU Xiao-Long[1], GAO Yuan-Yuan[1], ZHANG Kai-Hua[2]

[1](School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China)
[2](School of Computer Science, Nanjing University of Information Science & Technology, Nanjing 210044, China)

Abstract: Group activity recognition (GAR) is one of the highly researched areas in the field of computer vision, aiming to detect the overall behavior performed by multiple individual actions and interactions. However, due to difficulties in determining individual interaction relationships, the tightness of connections, and the key actor, current methods often focus on individual character features, yet neglecting connections with scene context. To address that issue, a novel reasoning model for GAR, GIFFNet, is proposed based on global-individual feature fusion (GIFF). To compensate for the lack of scene information in predicting group activity, GIFFNet, on the basis of focusing on key information, effectively integrates scene context and individual character features by constructing the GIFF module, obtaining more representative fusion features. Subsequently, GIFFNet utilizes fusion features to calculate the interaction relationship graph between characters in the scene and uses graph convolutional network (GCN) for training and predicting group behavior categories. In addition, to address the issue of imbalanced samples in the dataset, GIFFNet adopts a strategy of dynamically assigning weights to optimize the loss function. Experimental results demonstrate that GIFFNet achieves a

① 基金项目: 国家自然科学基金 (41975183, 41875184)
收稿时间: 2024-05-24; 修改时间: 2024-06-17; 采用时间: 2024-06-28; csa 在线出版时间: 2024-10-31

multi-class classification accuracy (MCA) of 93.8% and 96.1% on Volleyball and Collective Activity datasets, and the mean per class accuracy (MPCA) is 93.9% and 95.8%, respectively, outperforming other existing deep learning methods. GIFFNet provides features with a more powerful characterization ability for activity classification through feature fusion, which effectively improves GAR accuracy.

Key words: group activity recognition; scene context; feature fusion; attention mechanism; dynamic loss function

群体行为识别旨在理解和分析集体中多个个体的行为模式、交互及动态变化, 探索人群在不同情境下的行为特征和规律, 涉及机器学习、社会学和心理学等多个学科的交叉研究, 近年来得到广泛关注[1,2].

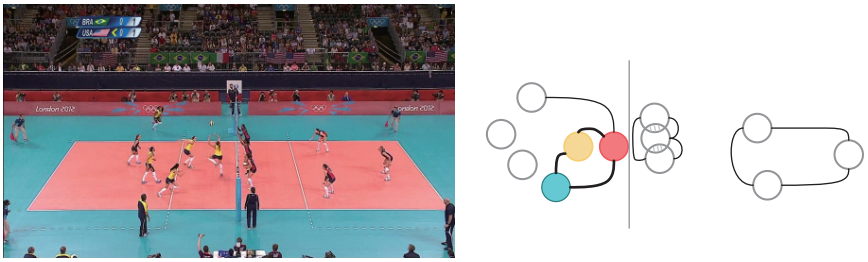
目前许多方法致力于构建多人场景下个体之间的互动关系, 以准确推断出群体行为. 早期基于循环神经网络 (RNN)[3−6]的方法可有效探索长时间段内的运动规律. 现有方法常使用卷积神经网络 (CNN)[7−9]与图卷积网络 (GNN)[10,11]模拟个体关系, 并结合注意力机制[12−15], 精细化提取人物的关系信息, 搭建关系交互图. 这些方法有效探索了集体中多个个体间的互动线索, 为群体行为识别的发展起到了促进作用.

尽管上述方法取得了良好的效果, 但仍存在以下问题.

(1) 忽略场景上下文信息, 致使行为识别精度较低. 在识别群体行为时, 场景上下文信息指人物所在场景中有助于判断个体与群体动作的环境和背景信息. 例如在排球运动中, 排球的运动轨迹通常会决定运动员的位置与行为倾向, 对运动员群体行为的判断起重要作用. 另外, 场景的布局与结构、其他人物的动态变化也会对目标人物的行为识别准确度产生一定的影响. 然而, 现有方法侧重于个体间的关系建模, 很少探索场景中与人物相关的线索. 这些方法通常直接从原始图像中提取包含了场景信息的全局特征, 然后基于标注的人物位置, 使用 RoI Align 对该特征进行感兴趣区域映射, 获取包含了人物细节信息的个体特征, 并最终利用个体特征预测群体行为的类别. 如图 1(a) 所示, 若仅利用标注框内的人物区域特征, 则会忽略其与全局场景的联系, 影响最终的行为识别效果.

(2) 图片中存在人物遮挡的情况, 致使缺失部分人物信息. 如图 1(a) 中右侧灰色框中的 3 位防守人物, 相互之间存在遮挡的情况, 无法精确地获取外观细节. 若仅利用其个体特征进行群体行为推理, 则会由于信息缺失而影响行为识别的精度.

(3) 难以确定关键人物, 致使分类结果被无关个体干扰. 场景中每个个体对群体行为的影响程度不同, 只有明确关键人物才能实现更加精确的群体行为分类. 如图 1(a)、(b) 所示, 其中标注的 3 位人物均具有较大的动作幅度, 相较于其他人物, 他们的互动关系也更为紧密, 对群体行为的识别结果起决定性作用. 若选择其他人物进行群体行为的判断, 则会使分类结果出错. 因此不能随意选取个体来判断整体行为或者给所有个体分配相同的参与权重, 否则会导致行为分类结果被无关个体所干扰. 如何充分利用人物所在场景的上下文信息, 并判断不同个体在识别过程中对群体行为的影响重要性, 是群体行为识别领域中亟待解决的重要问题. 此外, 在实际应用场景中, 数据集通常包含不同数量的样本类别, 导致模型倾向于预测出现频率更高的类别.



(a) 排球群体行为 (b) 人物联系

图 1 视频中排球群体行为的示例

为解决上述问题, 本文提出一个基于全局-个体特征融合的群体行为识别模型 GIFFNet (global-individual feature fusion network), 通过融合全局场景特征与个体人物特征, 解决行为识别时人物特征中场景信息缺失的问题. 具体而言, GIFFNet 使用 GIFF 模块将个体特征与全局特征紧密联合, 充分利用视频帧中的场景上下文信息. 通过将场景信息嵌入人物特征中, GIFFNet 不仅能够丰富特征训练时的场景信息, 还可以弥补人物遮挡的外观细节缺失问题. 其次, 若某些帧之间的场景变化较大, 可以通过融合全局特征捕捉这些动态变化, 提高模型的鲁棒性. 同时, 为了捕捉视频中的关键帧与图片中的关键人物, GIFF 模块中利用注意力机制聚焦关键信息, 实现精细化的特征提取. 另外, 针对图片分类的过程中数据集样本失衡的问题, 本文在标准交叉熵损失的基础上进行优化, 采用动态参数的思想, 取得显著的效果提升. 总体而言, 本文的贡献包括以下几个方面.

- 本文提出 GIFFNet, 融合了全局场景特征与个体人物特征, 并利用表征能力增强的融合特征进行群体行为的推理.
- 为高效整合场景上下文信息与个人特征, 本文提出了 GIFF 模块, 为个体动作与群体行为类别的推理提供了信息丰富的融合特征.
- 本文对损失函数进行优化, 采用动态分配权重的策略, 解决了数据集中潜在的样本失衡问题.
- 实验表明, 本文在两个大型的群体行为数据集——Volleyball[3]与 Collective Activity[16]上, 取得了出色的群体行为识别准确度.

1 相关工作

群体行为分析与识别是计算机视觉和人工智能领域中的重要研究方向之一, 其目标是通过对视频数据的准确分析, 捕捉群体中个体之间的互动和协同行为. Choi 等人[16]提出群体行为识别的概念, 随后研究者们展开一系列研究.

1.1 基于传统深度学习的群体行为识别方法

早期研究者们使用传统方法和特征工程实现群体行为识别. 该类方法根据视频帧中人员的运动轨迹, 结合目标检测和目标跟踪等技术, 分析运动者的空间位置信息以确定动作的类别. 然而, 由于人物外观信息以及场景信息的重要性, 仅使用空间位置信息难以实现高精度的识别结果. 为解决这一问题, 研究者们引入一些手工设计的特征描述子, 如早期广受关注的方向梯度直方图 (HOG) 特征描述子. 这些特征描述子能够提取人物的空间信息、运动信息、场景上下文等, 以更加准确地判断人物动作.

随着深度学习的兴起, 其在群体行为识别领域内取得了巨大成功. Ibrahim 等人[3]提出一个层次化的时序深度学习模型, 利用长短时记忆网络 (LSTM) 在长时间段内建模人物的动态信息, 并结合卷积网络提取场景的全局特征, 实现对群体行为的分类. 此外, 作者提出了 Volleyball 数据集, 作为一个用于群体行为识别方向的大型数据集, 为该领域的后续研究做出了巨大贡献. 意识到人与人之间互动关系在群体行为识别中的重要性后, Wu 等人[10]提出构建关系图的方法, 基于图的数据结构, 有效地在视频帧中的人员之间搭建起了联系. Yuan 等人[11]在此基础上, 提出时空双线性池化的方法, 进一步探索关系图的结构. 同年, Yuan 等人[8]又结合人物的时空交互关系, 通过构建关系域, 动态地探索人物之间的联系. 为解决数据集标签错误的问题, Demirel 等人[12]尝试重新标注 Volleyball 数据集中错误的样本, 并通过注意力池化有效提升了群体行为识别的准确度. 此外, 为解决数据集标签缺失的问题, 半监督与无监督学习的方法[17–19]也逐渐应用于群体行为识别领域中, 并取得了显著成果. 与上述方法不同的是, 本文致力于将场景特征与个体特征进行融合, 并利用融合后的增强特征进行动作推理, 以探求更好的分类效果.

1.2 基于 Transformer 的群体行为识别方法

2017 年, Vaswani 等人[20]首次提出注意力机制的概念与 Transformer 模型. 这一机制模拟人类视觉系统的工作方式, 针对不同的输入部分分配不同的注意力权重, 实现对重要信息的聚焦. 受注意力机制思想的启发, 后续研究者们进行创新, 诞生了一系列性能优秀的算法, 如 SENet[21]等.

最初, 注意力机制应用于语义分割的领域. 为解决图像识别问题, Dosovitskiy 等人[22]提出 Vision Transformer (ViT) 模型. ViT 首先将整个图像分割成 16 个较小的 patch, 然后对每个 patch 进行线性映射与位置编码, 以获取 ViT 编码器的输入. ViT 编码器与原始的 Transformer 相同, 使用了多头自注意力 (MSA) 机制, 以便聚焦图像中的重要信息, 并通过 MLP 层获取分类结果. 整个模型可包含多层, 根据训练效率和分类结果确定最终的层数. Liu 等人[23]提出 Swin Transformer, 利用 SW-MSA 机制偏移图片像素并窗口化输入特征, 在

降低计算量的同时, 实现更高的识别准确度, 成为 Transformer 中的经典模型之一. 为了将更加轻量化的 ViT 模型应用于计算机视觉领域, Mehta 等人[24,25]提出 MoblieViT 模型, Vasu 等人[26]提出 FastViT 模型, 在保持原有识别效率的基础上, 减少了资源消耗. 在群体行为识别领域中, 结合 Transformer 等注意力机制模型, 可对提取的人物特征进行细化. Han 等人[27]提出一个基于时空双路径 Transformer 的群体行为分类模型, 通过组合时-空与空-时两个不同顺序的网络, 多角度地分析人物动作特征. Li 等人[28]将 Transformer 与聚类结合, 提出用于群体行为识别的模型 GroupFormer. Tamura 等人[29]针对社会群体行为活动, 利用 Transformer 模型细分子类群体行为活动, 探索子类群体之间的关系线索. 然而, 这些方法未能提前融合人物细节与场景特征, 导致在预测群体行为时缺失场景信息. 为了解决这一问题并捕捉人物所在场景中的重要信息, 本文提出了 GIFF 模块, 将场景信息嵌入人物特征中, 为群体行为推理提供了良好的条件. GIFF 模块借鉴了 Transformer 模型的结构, 能够有效增强特征的表征能力.

2 方 法

在本节中, 首先对模块的整体框架做出概述, 然后针对框架中的子模块进行详细介绍. 相对于现有方法, GIFFNet 不仅能够细致捕获人物互动关系, 还强调了结合场景上下文信息的重要性. 接着, 对框架内部具体模块展开阐述, 包括全局-个体特征融合 (GIFF) 模块、融合时机的评估以及损失函数的优化等.

2.1 网络概述

本文提出的网络分为两个主要阶段. 在第 1 阶段中, 首先, 选取数据集中的图像, 利用骨干网络提取图像原始特征, 即包含了场景信息的全局特征. 接着, 结合人物在场景中的位置, 使用 RoI Align 对全局特征进行感兴趣区域映射, 得到包含人物外观细节的个体特征. 全局特征与个体特征是第 1 阶段的输出, 为特征融合与群体行为预测的基础. 而第 2 阶段主要负责特征融合与群体行为的预测. 为实现个体特征与全局特征的融合, GIFFNet 构建了 GIFF 模块, 将场景信息嵌入个体特征中, 获取能够同时表达场景信息与人物细节的融合特征, 解决了场景信息缺失与人物遮挡等问题, 为识别群体行为提供了良好的条件. 在确定特征的融合时机并进行特征融合后, 计算人物之间的联系以构建关系交互图. 最终, 利用图卷积网络对关系交互图进行训练, 并使用分类器预测群体行为的标签, 计算群体行为的识别准确度与训练损失. 本文采用全监督的方法, 模型整体框架如图 2 所示, 两个阶段的具体流程如下.

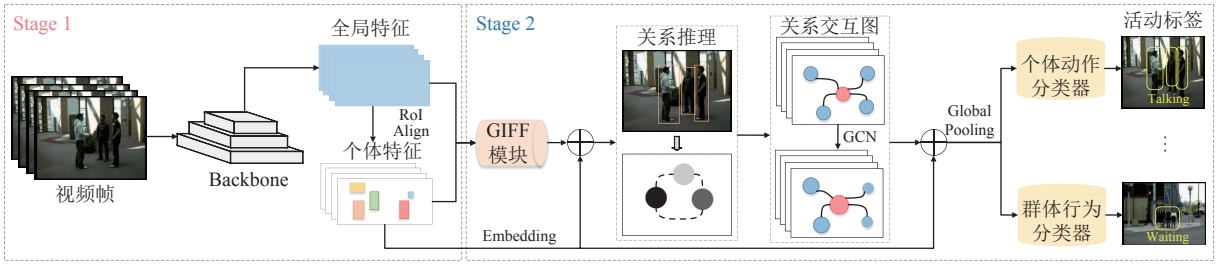


图 2 GIFFNet 整体框架

Stage 1: 将原始的视频裁剪成帧, 以 RGB 模态的图像为输入, 利用骨干网络分别提取图像的全局特征 $\boldsymbol{X}_g \in R^{BT\times C_g\times W\times H}$, 其中 B 代表批量大小, T 代表时间步长, C_g 代表全局图像特征的通道维度, W 与 H 代表图像的尺寸; 接着, 结合全局特征 $\boldsymbol{X}_g$ 与标注的人物位置, 采用 RoI Align 映射来提取包含外观细节的个体人物特征 $\boldsymbol{X}_i \in R^{BTN\times C_i\times K\times K}$, 其中 C_i 代表个体特征的通道维度, N 代表目标人物数量, K 代表裁剪后的图像尺寸.

Stage 2: 将 Stage 1 中得到的全局特征 $\boldsymbol{X}_g$ 和个体特征 $\boldsymbol{X}_i$ 送入 Stage 2 中的网络进一步训练, 实现个体人物动作的识别, 并据此进一步推理群体行为类别. 首先, 通过 GIFF 模块将特征 $\boldsymbol{X}_g$ 与 $\boldsymbol{X}_i$ 融合, 将 $\boldsymbol{X}_g$ 中包含的场景上下文信息融入 $\boldsymbol{X}_i$ 中, 输出增强后的个体特征 $\boldsymbol{X}_{\text{fused}} \in R^{B\times C_i\times T\times N}$. 其中, 个体特征在经过嵌入层后, 与 GIFF 模块的输出进行残差计算, 获得最终的融合特征 $\boldsymbol{F}_{\text{fused}} \in R^{B\times T\times N\times C_i}$. 然后, 利用 $\boldsymbol{F}_{\text{fused}}$ 计算人物之间的关系交互矩阵, 以推断群体行为的类别. 具体来说, 首先需计算 $\boldsymbol{F}_{\text{fused}} = \{\boldsymbol{x}_n\}_{n=1}^{TN}$ 中不同人物特征之间的外观相似度与位置差异, 以获取人物之间的关系交互图.

$$\boldsymbol{G}_{i,j} = h(f_a(\boldsymbol{x}_i^a, \boldsymbol{x}_j^a), f_s(\boldsymbol{x}_i^s, \boldsymbol{x}_j^s)) \tag{1}$$

其中, $\boldsymbol{G}_{i,j}$ 为人物 i 与 j 之间的关系交互图, $f_a(\boldsymbol{x}_i^a, \boldsymbol{x}_j^a)$ 与

$f_s(\boldsymbol{x}_i^s,\boldsymbol{x}_j^s)$分别代表两个人物之间的外观相似度与位置关系. f_a 与 f_s 为计算外观相似度与位置差异的函数, h 为融合二者的函数. 接着, 需利用图卷积网络 (GCN) 对该特征图进行训练.

$$\boldsymbol{X}^{l+1}=\sigma(\boldsymbol{G}\boldsymbol{X}^l\boldsymbol{W}^l)+\boldsymbol{X}^l \tag{2}$$

其中, $\boldsymbol{G}$ 代表人物之间的关系交互图, $\boldsymbol{X}^l$ 代表第 l 层的人物特征表示, 其中 $\boldsymbol{X}^0=\boldsymbol{F}_{\text{fused}}$. $\boldsymbol{W}^l$ 为第 l 层可学习的权重矩阵, σ为 *ReLU* 激活函数. 根据式 (2), GCN 的传播训练可以设计为多层. 为避免较大的计算复杂度, 本文仅使用一层 GCN 进行特征训练, 因此式 (2) 中 $\boldsymbol{X}^l$ 即为 $\boldsymbol{F}_{\text{fused}}$.

接着, 将 GCN 训练得到的输出特征与原始的个体特征相加, 经过全局平均池化层, 并通过个体动作分类器与群体行为分类器, 以获取最终的个体动作标签与群体行为标签. 群体行为标签的预测准确度是本文评判模型的重要指标. 另外, 为解决数据集样本失衡的问题, 本文在标准交叉熵损失的基础上采用动态分配权重的策略进行改进.

2.2 全局-个体特征融合 (GIFF) 模块

GIFF 整体框架如图 3 所示.

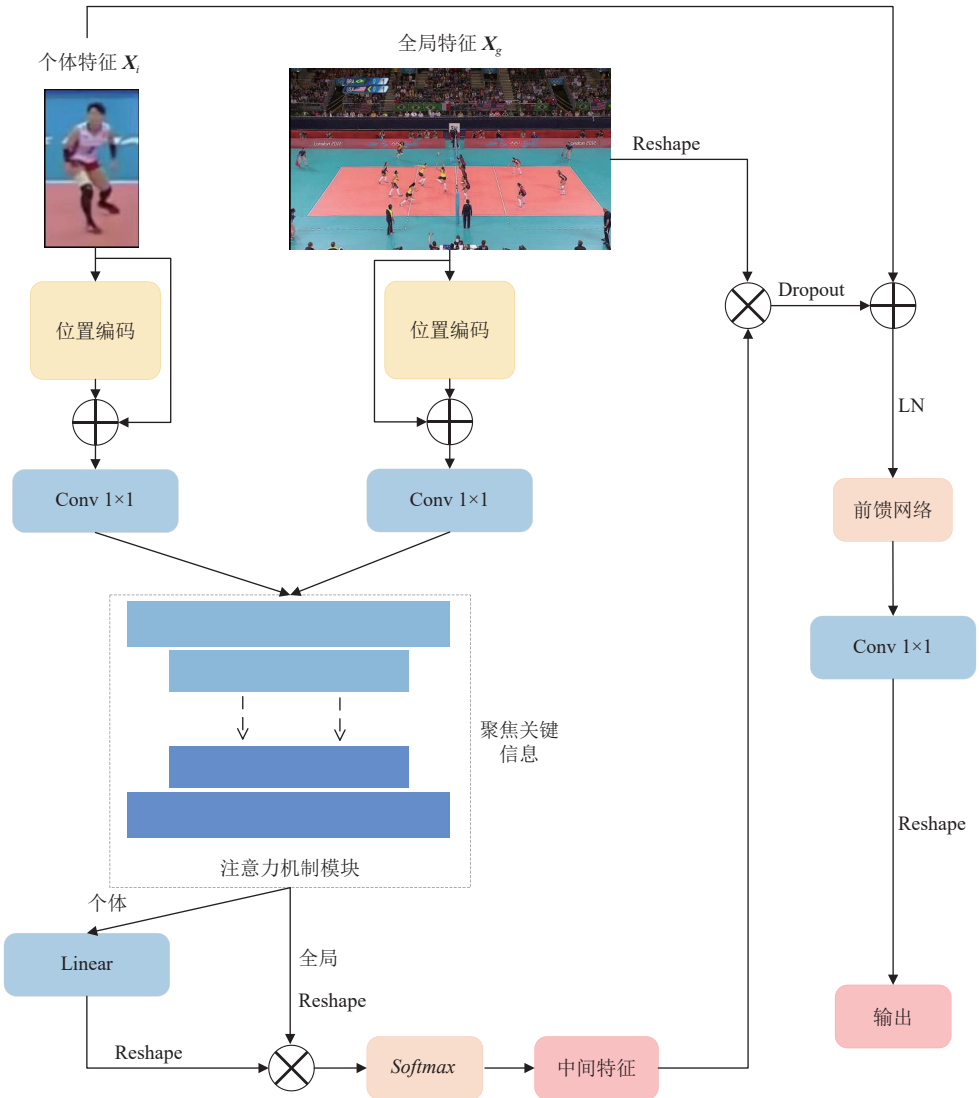


图 3 GIFF 模块整体框架

在传统的群体行为识别算法中, 通常直接对个体特征进行学习训练, 随后进行群体行为的预测. 然而, 这样的做法仅利用了人物的细节信息, 忽略了人物与场景信息之间的重要关联. 例如, 在排球运动中, 球的

位置与运动轨迹常常影响了运动员的行为执行策略. 如果仅利用运动员的姿态、外观等细节信息, 而忽略场景中球与运动员的联系, 难以进一步判断人物的行为类别. 此外, 如果人物被场景中的物体或其他人物遮挡, 则会导致预测群体行为时缺失某些重要的人物细节信息. 为弥补这些缺陷, 本文提出 GIFF 模块, 用于融合个体特征与全局特征, 将场景上下文信息嵌入个体特征中, 以增强特征的表达能力.

GIFF 模块的输入包括两部分: 第 1 阶段输出的个体特征 $\boldsymbol{X}_i \in R^{BTN\times C_i\times K\times K}$ 与全局特征 $\boldsymbol{X}_g \in R^{BT\times C_g\times W\times H}$. 针对性地将个体特征与场景上下文进行融合, 生成富含场景信息的个体特征. 该模块能够为群体行为识别提供更具表征能力的特征, 实现高准确度的群体行为分类结果. GIFF 模块的伪代码如算法 1 所示, 其具体流程如下.

算法 1. GIFF 模块

输入: 全局特征 $\boldsymbol{X}_g$, 个体特征 $\boldsymbol{X}_i$.
输出: 融合特征 $\boldsymbol{X}_{\text{fused}}$.

1. 将 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_g$ 输入 GIFF 模块;
2. 位置编码与下采样: 利用式 (3) 对 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_g$ 进行位置编码与下采样处理, 得到 $\boldsymbol{X}'_i$ 与 $\boldsymbol{X}'_g$;
3. 注意力机制模块: 利用式 (4) 对 $\boldsymbol{X}'_i$ 与 $\boldsymbol{X}'_g$ 应用注意力机制, 得到 $\widehat{\boldsymbol{X}}'_i$ 与 $\widehat{\boldsymbol{X}}'_g$;
4. $\widehat{\boldsymbol{X}}'_i$ 经过全连接层得到 $\widehat{\boldsymbol{X}}'_{i\_emb}$;
5. 调整 $\widehat{\boldsymbol{X}}'_g$ 与 $\widehat{\boldsymbol{X}}'_{i\_emb}$ 的张量形状;
6. 乘积融合与归一化: 利用式 (6) 对 $\widehat{\boldsymbol{X}}'_g$ 与 $\widehat{\boldsymbol{X}}'_{i\_emb}$ 进行融合, 得到中间特征 $\boldsymbol{X}_{\text{tmp}}$;
7. 乘积融合与残差计算: 利用式 (7) 将 $\boldsymbol{X}_{\text{tmp}}$ 与 $\widehat{\boldsymbol{X}}'_g$ 乘积融合并与 $\widehat{\boldsymbol{X}}'_{i\_emb}$ 进行残差计算, 生成融合了场景上下文信息的人物特征 $\widehat{\boldsymbol{X}}''_i$;
8. 正则化与前馈网络: 利用式 (8) 对 $\widehat{\boldsymbol{X}}''_i$ 进行正则化处理, 并通过 FFN, 得到 $\widehat{\boldsymbol{X}}_{\text{fused}}$;
9. 利用式 (9) 调整 $\widehat{\boldsymbol{X}}_{\text{fused}}$ 的张量形状, 得到融合特征 $\boldsymbol{X}_{\text{fused}}$;
10. 输出 $\boldsymbol{X}_{\text{fused}}$;

首先, 为了更好地理解图像中不同区域的重要性, GIFF 模块利用位置编码 (position encoding, PE) 处理 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_g$, 为其添加位置信息. 此外, 为了减少参数量, GIFF 模块采用了 bottleneck 结构, 通过使用 1×1 卷积核的 pointwise 卷积层来统一规划两种特征的通道数.

$$\boldsymbol{X}' = Conv_{1\times1}(PE(\boldsymbol{X}) + \boldsymbol{X}) \tag{3}$$

其中, PE 与 $Conv_{1\times1}$ 分别代表位置编码和 pointwise 卷积层. $\boldsymbol{X}$ 表示个体特征 $\boldsymbol{X}_i$ 和全局特征 $\boldsymbol{X}_g$, $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_g$ 经过式 (3) 运算后输出 $\boldsymbol{X}'_i$, 包括 $\boldsymbol{X}'_i \in R^{BTN\times C_d\times K\times K}$ 与 $\boldsymbol{X}'_g \in R^{BT\times C_d\times W\times H}$, C_d 为统一后的通道数.

为进一步提炼两种特征的关键信息, GIFF 模块添加了通道注意力机制, 细化了 $\boldsymbol{X}'_i$ 与 $\boldsymbol{X}'_g$ 的通道, 以捕获视频序列中的关键帧并聚焦群体中的关键人物.

$$\widehat{\boldsymbol{X}}' = ChannelAtt(\boldsymbol{X}') = w_{b\times c\times1\times1} \times \boldsymbol{X}' \tag{4}$$

其中, w 是注意力模块中训练所得的权重矩阵, 与特征 $\boldsymbol{X}'_i$ 和 $\boldsymbol{X}'_g$ 在前两个维度上相乘, 根据 $\boldsymbol{X}'_i$ 和 $\boldsymbol{X}'_g$ 中特征的重要性进行加权运算得到 $\widehat{\boldsymbol{X}}'_i$ 与 $\widehat{\boldsymbol{X}}'_g$.

接着, $\widehat{\boldsymbol{X}}'_i$ 将通过一个线性嵌入层, 降低维度并调整形状得到 $\widehat{\boldsymbol{X}}'_{i\_emb} \in R^{B\times C_d\times T\times N}$. 随后, 将两种特征进行乘积融合, 并通过 *Softmax* 层得到中间特征 $\boldsymbol{X}_{\text{tmp}} \in R^{BT\times N\times WH}$.

$$\widehat{\boldsymbol{X}}'_{i\_emb} = Linear(\widehat{\boldsymbol{X}}'_i) \tag{5}$$

$$\boldsymbol{X}_{\text{tmp}} = Softmax(\widehat{\boldsymbol{X}}'_g \times \widehat{\boldsymbol{X}}'_{i\_emb}) \tag{6}$$

其中, $Linear$ 与 $Softmax$ 分别表示线性嵌入层与 *Softmax* 层.

GIFF 模块将 $\boldsymbol{X}_{\text{tmp}}$ 与 $\widehat{\boldsymbol{X}}'_g$ 乘积融合, 生成包含全局场景信息的特征. 然后, 将该特征与 $\widehat{\boldsymbol{X}}'_{i\_emb}$ 进行残差计算, 保留个体特征中的原有信息, 并生成融合了场景上下文信息的特征 $\widehat{\boldsymbol{X}}''_i$.

$$\widehat{\boldsymbol{X}}''_i = \boldsymbol{X}_{\text{tmp}} \times \widehat{\boldsymbol{X}}'_g + \widehat{\boldsymbol{X}}'_{i\_emb} \tag{7}$$

$\widehat{\boldsymbol{X}}''_i$ 需经过层归一化与前馈网络, 通过对 $\widehat{\boldsymbol{X}}''_i$ 进行复杂的非线性变换与正则化操作, 进一步增强其表达能力.

$$\widehat{\boldsymbol{X}}_{\text{fused}} = LN(FFN(LN(\widehat{\boldsymbol{X}}''_i))) \tag{8}$$

其中, LN 和 FFN 分别代表层归一化和前馈网络, $\widehat{\boldsymbol{X}}_{\text{fused}} \in R^{B\times C_d\times T\times N}$ 为表征能力增强的融合特征.

如图 2 所示, GIFF 模块的输出需与 $\boldsymbol{X}_i$ 再次进行残差相加. 为了便于这两个特征的运算, GIFF 模块内使用了一个 1×1 大小卷积核的 pointwise 卷积层, 对 $\boldsymbol{X}_{\text{fused}}$ 的通道数进行还原, 并调整 $\boldsymbol{X}_{\text{fused}}$ 使其与 $\boldsymbol{X}_i$ 的形状保持一致.

$$\boldsymbol{X}_{\text{fused}} = Conv_{1\times1}(\widehat{\boldsymbol{X}}_{\text{fused}}) \tag{9}$$

其中, $\boldsymbol{X}_{\text{fused}} \in R^{B\times T\times N\times C_i}$ 为 GIFF 模块最终输出的融合特征.

总体而言, GIFF 模块主要起到了以下两个作用.

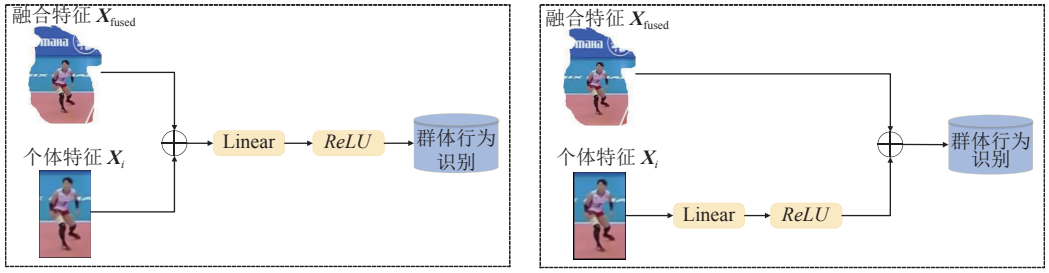
首先, GIFF 模块对全局特征与个体特征进行位置编码与网络层训练, 使其带有位置信息并提取有效的特征表示. 然后, 将这两类特征通过乘积融合与残差融

合的方式进行结合. 乘积融合能够捕捉它们之间的细微关系, 而残差融合则保留了原始特征信息. 这一过程将全局特征中包含的场景信息融入个体特征, 使融合后的特征能够同时表达人物的细节信息与场景信息, 从而为群体行为识别提供了充分的条件.

其次, GIFF 模块内的注意力机制模块动态地为两类特征分配权重, 能够聚焦于视频中的关键人物和数据集中的关键帧. 这一机制有效地防止了无关因素对后续群体行为识别流程的干扰, 提高了模型的识别准确性和鲁棒性. 另外, 考虑到运算复杂度大的问题, GIFF 模块借鉴了 bottleneck 的结构, 有效地降低了参数量.

2.2 全局-个体特征融合 (GIFF) 模块

GIFF 模块成功地将场景信息嵌入个体特征中, 但为了更充分地保留个体特征中的人物细节信息, GIFF 模块的输出 $\boldsymbol{X}_{\text{fused}}$ 需要与个体特征 $\boldsymbol{X}_i$ 再做一次残差计算. 针对这一过程, 本文提出了两种不同的融合方式, 根据融合时机的差异, 分为早期融合和晚期融合, 如图 4 所示.



(a) 早期融合 (b) 晚期融合

图 4 两种特征融合方式

图 4(a) 展现了早期融合策略, GIFF 模块输出 $\boldsymbol{X}_{\text{fused}}$ 后直接与 $\boldsymbol{X}_i$ 进行相加融合, 然后经过线性嵌入层提炼特征信息, 并利用 *ReLU* 激活函数进行非线性变换, 得到最终的融合特征 $\boldsymbol{F}_{\text{fused}}$. 早期融合策略简化了融合步骤, 能够快速融合特征并进行后续处理, 但可能会导致部分原始特征信息在融合过程中被忽略. 图 4(b) 展现了晚期融合策略, $\boldsymbol{X}_i$ 首先单独经过线性嵌入层与 *ReLU* 激活函数处理, 随后再与 $\boldsymbol{X}_{\text{fused}}$ 相加融合, 获取最终的融合特征 $\boldsymbol{F}_{\text{fused}}$. 晚期融合策略保留了 $\boldsymbol{X}_i$ 中更多的原始特征信息, 通过优先对 $\boldsymbol{X}_i$ 进行处理, 可以更充分地提取目标人物的细节信息, 使融合后的特征更具表征能力. 尽管该策略的融合步骤相对复杂, 但能够有效提升特征的表示能力. 通过对这两种策略的分析与比较, 可以发现早期融合策略适用于需要快速处理和融合特征的场景, 而晚期融合策略则更适用于需要保留更多原始人物特征信息并进行深入特征提取的场景. 式 (10) 和式 (11) 展现了早期融合与晚期融合的计算过程.

$$\boldsymbol{F}^{e}_{\text{fused}} = ReLU(LN(Linear(\boldsymbol{X}_{\text{fused}} + \boldsymbol{X}_i))) \tag{10}$$

$$\boldsymbol{F}^{l}_{\text{fused}} = ReLU(LN(Linear(\boldsymbol{X}_i))) + \boldsymbol{X}_{\text{fused}} \tag{11}$$

其中, $\boldsymbol{F}^{e}_{\text{fused}}$ 与 $\boldsymbol{F}^{l}_{\text{fused}}$ 分别是通过早期融合和晚期融合得到的融合特征, 该特征将用于后续群体行为的预测.

2.4 损失函数

现有群体行为识别方法通常使用标准交叉熵损失来确定预测动作与真实动作之间的差距, 但并未解决样本分类不平衡的问题. 当数据集中某类样本量远超其他样本或数据集样本量较小时, 将会导致模型倾向于预测占比量更多的类别, 从而影响模型的性能. 例如在 Collective Activity 数据集中, 样本数据量仅有约 1G 大小, 其较小的样本量会导致分类样本的失衡. 通过调整损失函数, 可以提升模型对少数类别的预测性能, 在一定程度上缓解样本失衡问题. 为解决该问题, 本文在标准交叉熵损失的基础上, 采取动态分配权重的策略, 针对损失函数做出了一定的改进, 对最终的行为分类起到了较好的促进效果. 具体介绍如下.

$$loss = \alpha(1-pt)^{\gamma} L(\boldsymbol{y}_a, \bar{\boldsymbol{y}}_a) \tag{12}$$

$$pt = \exp(L(\boldsymbol{y}_a, \bar{\boldsymbol{y}}_a)) \tag{13}$$

其中, L 代表交叉熵损失函数, $\boldsymbol{y}_a$ 和 $\bar{\boldsymbol{y}}_a$ 分别是真实标签值和预测标签值, α 和 γ 是超参数. 具体而言, 在计算得到预测动作与真实标签之间的交叉熵损失后, 需利用式 (12) 进一步计算损失大小. α 和 γ 的值可依据实际需求与最终效果自由调整, 参数 pt 利用式 (13) 计算获得.

3 实验分析

3.1 数据集

为应对群体行为识别领域中的多种挑战, 研究者们开发并广泛应用两个经典数据集, 分别为 volleyball dataset (VD) 和 collective activity dataset (CAD).

Volleyball 数据集是一个用于群体行为识别和分析的数据集, 专注于排球比赛中的群体动作设计. 该数据集旨在帮助研究人员开发算法来识别和分析排球比赛中的不同群体行为, 如发球、传球、扣球、拦网等. 其在 55 场大型排球赛事中进行动作的采集, 包括了 4 830 个视频片段. 按照排球比赛的动作术语, 定义了 8 个群体行为标签: right set、right spike、right pass、right winpoint、left set、left spike、left pass、left winpoint 和 9 个个体行为标签: waiting、setting、digging、failing、spiking、blocking、jumping、moving、standing.

Collective Activity 数据集涵盖 44 个视频片段, 每个视频片段约有 200–1800 个视频帧. 该数据集采用监控摄像头进行视频采集, 旨在捕捉多个行动者在不同场景中的集体活动. 根据场景中行动者的动作差异, 设置了 5 个群体行为标签: crossing、waiting、queueing、walking、talking 和 6 个个体行为标签: NA、crossing、waiting、queueing、walking、talking.

3.2 实验细节

本文在一张 NVIDIA A800 (80 GB) 的显卡上对两个数据集进行实验, 整个模型使用 PyTorch 框架实现. 为了减少消耗并提高代码运行性能, 实验中使用了预训练的模型结果, 据此进行进一步的模型拟合与推理判断. 在两个群体数据集上, 本文设置约 70% 的数据为训练集, 30% 的数据为测试集. 在 Volleyball 数据集中, 输入图像的大小为 (720, 1024), 选取的骨干网络为 VGG-16, 共设置了 60 轮的训练与测试, 初始学习率 lr 设为 1^{-4}, 每 10 轮缩减 1/2, 并在第 40 轮固定为 1^{-5}; 在 Collective Activity 数据集中, 输入图像的大小为 (480, 720), 选取的骨干网络为 Inception-v3, 共设置了 30 轮的训练与测试, 学习率 lr 固定为 5^{-5}. 另外, 在两个数据集上训练与测试时, dropout ratio 设为 0.3, 权重衰退设为 1^{-4}, 损失函数计算中 α 和 γ 的值分别为 0.25 和 0.2. 衡量准确度的两个指标分别为多类分类准确度 (multi-class classification accuracy, MCA) 和类平均精确度 (mean per class accuracy, MPCA).

3.3 对比实验

表 1 展示了 GIFFNet 与其他流行算法在 Volleyball 数据集上群体行为的识别准确度对比结果. 在该数据集的实验中, GIFFNet 选用的骨干网络为 VGG-16. 结果显示 GIFFNet 的 MCA 值与 MPCA 值分别高达 93.8% 与 93.9%, 个体动作的准确率也达到 83.6%, 优于多个经典的群体行为识别算法, 甚至在准确度上超越近年来提出的高精度识别模型.

表 1 本文方法与其他流行方法在 Volleyball 数据集上的识别率对比 (%)

Methods	Backbone	MCA	MPCA	Action Acc
HDTM[3]	AlexNet	81.9	82.9	—
SSU[30]	Inception-v3	89.9	—	81.8
PCTDM[31]	ResNet-18	90.3	90.5	—
stagNet[6]	VGG-16	89.3	—	—
CRM[9]	I3D	92.1	—	—
ARG[10]	ResNet-18	91.1	91.4	83.0
PRL[32]	VGG-16	91.4	91.8	—
STBiP[11]	Inception-v3	91.3	—	—
SACRF[14]	ResNet-18	90.7	91.0	83.1
DIN[7]	ResNet-18	93.1	93.3	—
GroupFormer[28]	Inception-v3	93.4	—	83.2
TSG+PDFM[33]	Inception-v3	92.4	—	—
Ours-GIFFNet	VGG-16	**93.8**	**93.9**	**83.6**

注: 加粗字体表示最优结果, “—”表示准确率未提供

表 2 展示了 GIFFNet 与其他流行算法在 Collective Activity 数据集上群体行为的识别准确度对比结果. 在该数据集的实验中, GIFFNet 选用的骨干网络为 Inception-v3. 比较来看, GIFFNet 在样本数量较少的数据集上同样取得出色的识别效果, 模型精度优于几年提出的流行算法, 其 MCA 和 MPCA 值分别高达 96.1% 与 95.8%.

表 2 本文方法与其他流行方法在 Collective Activity 数据集上的识别率对比 (%)

Methods	Backbone	MCA	MPCA
HDTM[3]	AlexNet	90.4	89.7
CERN[34]	VGG-16	87.2	88.3
stagNet[6]	VGG-16	87.7	89.1
PCTDM[31]	AlexNet	92.1	92.2
ARG[10]	ResNet-18	92.4	92.3
PRL[32]	VGG-16	—	93.8
HiGCIN[35]	ResNet-18	93.4	93.0
DIN[8]	ResNet-18	—	95.3
GroupFormer[28]	Inception-v3	93.6	—
TSG+PDFM[33]	Inception-v3	93.5	—
DECOMPL[12]	VGG-16	95.5	—
Ours-GIFFNet	Inception-v3	**96.1**	**95.8**

从两个测试数据集上的实验结果来看, GIFFNet 通常比基于 RNN 的算法具有更优秀的性能, 同时相对开销更小. 相较于基于 GNN 与 CNN 的方法, GIFFNet 也取得精度上的优势. 在样本量较小的 Collective Activity 数据集上, GIFFNet 实现更高的识别准确度, 优于其他的流行算法.

3.4 消融实验

为验证特征融合模块和损失函数对整体网络的有效性与兼容性, 本文针对这两个方面实施详尽的消融实验. 表 3 与表 4 分别是在两个群体行为数据集上展开的消融实验数据, 结果表明本文提出的融合模块与改进的损失函数均对识别结果有显著提升.

表 3 针对融合模块和损失函数在 Volleyball 数据集上的消融数据 (%)

Model	MCA	MPCA
Base model	93.2	93.2
Base model + GIFF module	93.6	93.6
Base model + loss function optimization	93.4	93.5
GIFFNet	**93.8**	**93.9**

表 4 针对融合模块和损失函数在 Collective Activity 数据集上的消融数据 (%)

Model	MCA
Base model	95.3
Base model + GIFF module	95.7
Base model + loss function optimization	95.8
GIFFNet	**96.1**

本文基础模型使用的损失函数为标准交叉熵损失, 根据 Volleyball 数据集上的消融实验结果, 其准确率约为 93.2%. 在基础模型的基础上, 单独添加 GIFF 模块后, 能够将全局场景信息嵌入个体特征中并加以预测, 其准确度分别提升至 93.6%. 其原因是通过融合全局场景信息后, 能够有效弥补信息缺失的问题, 以提高模型的识别准确度. 另外, 在标准交叉熵损失函数的基础上, 单独改进损失函数后, 一定程度上解决了样本失衡的问题, 使识别准确度提升至 93.4%. 将损失函数改进并添加 GIFF 模块后, 模型的识别准确率进一步提高, 高达 93.8%. 这表明 GIFFNet 在 Volleyball 数据集上取得了较好的性能提升效果. 在 Collective Activity 数据集的消融实验中, 基础模型的准确率约为 95.3%, 其损失函数同样为标准交叉熵损失. 随后, 单独添加 GIFF 模块后, 准确率提升至 95.7%. 仅对损失函数进行改进的情况下, 准确率达到 95.8%. 最后, 同时添加 GIFF 模块并对损失函数进行改进时, 识别准确率显著提高至 96.1%. GIFFNet 在 Collective Activity 数据集上取得了显著的性能提升, 证明了以融合场景信息的个体特征为识别模型的输入能够提高整体网络的鲁棒性.

为验证 GIFF 模块的最佳适用性, 本文在两个数据集上展开消融实验. 根据融合时机的差异, 对 GIFF 模块的输出与个体特征进行外部融合, 实验结果如表 5 与表 6 所示. 在早期融合中, GIFF 模块的输出特征 $\boldsymbol{X}_{\text{fused}}$ 直接与原个体特征 $\boldsymbol{X}_i$ 相加, 再经过嵌入层以得到最终的融合特征, 这可能会导致 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_{\text{fused}}$ 中部分特征的信息丢失与混合. 而在晚期融合中, 在个体特征 $\boldsymbol{X}_i$ 获取嵌入信息后, 再与 GIFF 模块的输出特征 $\boldsymbol{X}_{\text{fused}}$ 相加获取融合特征, 能够使模型更充分地利用输入特征 $\boldsymbol{X}_i$ 中的信息, 与 $\boldsymbol{X}_{\text{fused}}$ 进行更有效的融合, 使得最终的融合特征能够更好地表示人物之间的关系. 实验结果显示, 结合晚期融合的模型识别准确率更高, 在 Volleyball 数据集上, 早期融合和晚期融合的准确率分别为 93.2% 和 93.8%; 而在 Collective Activity 数据集上, 准确率分别为 95.8% 和 96.1%. 由于晚期融合能够更充分地利用输入特征的信息, 减少信息丢失, 因此相较于早期融合, 其效果更佳.

表 5 针对融合时机在 Volleyball 数据集上的消融数据 (%)

Fusion methods	MCA
Early fusion	93.2
Late fusion	**93.8**

表 6 针对融合时机在 Collective Activity 数据集上的消融数据 (%)

Fusion methods	MCA
Early fusion	95.8
Late fusion	**96.1**

对图像进行位置编码能够更好地理解图像特征上的位置信息. 为验证图像编码在本文模型中的有效性, 本文在 Volleyball 数据集上展开消融实验, 实验结果如表 7 所示. 实验结果显示, 在 GIFF 模块中不包含位置编码的情况下, 准确率仅有 93.2%; 当对 GIFF 模块中的全局特征进行位置编码后, 准确率提升至 93.5%; 最后, 对 GIFF 模块中的两类特征都进行位置编码, 准确率进一步提升至 93.8%. 这表明位置编码对模型理解图像特征起到积极的促进作用.

表 7 针对位置编码在 Volleyball 数据集上的消融数据 (%)

Model	MCA
GIFF module wo/ PE	93.2
Global Features of GIFF module w/ PE	93.5
GIFF module w/ PE	**93.8**

3.5 可视化

图 5 和图 6 展示了 GIFFNet 在两个数据集上混淆矩阵的可视化结果, 均取得了出色的识别准确率. 对于 Volleyball 数据集中的每个动作, GIFFNet 的识别准确率都超过 90%. 在 Collective Activity 数据集的动作识别中, 为了简化模型的输出的复杂度, 将行为较为相近的 moving 和 crossing 标签映射为一个类别 moving/crossing, 使群体活动的类别从 5 个简化为 4 个. 结果显示 moving/crossing 和 waiting 两类行为的识别存在一定误差, 这是由于在某些场景下, 两者的动作形态较为相似, 同时数据集样本较少也会对最终的分类结果产生一定的影响.

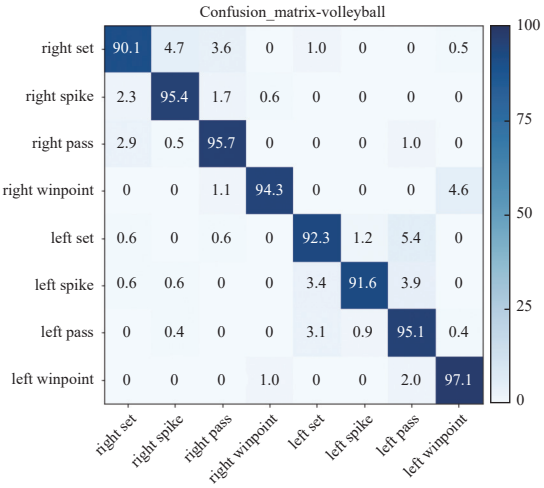


图 5　GIFFNet 在 Volleyball 数据集上的混淆矩阵

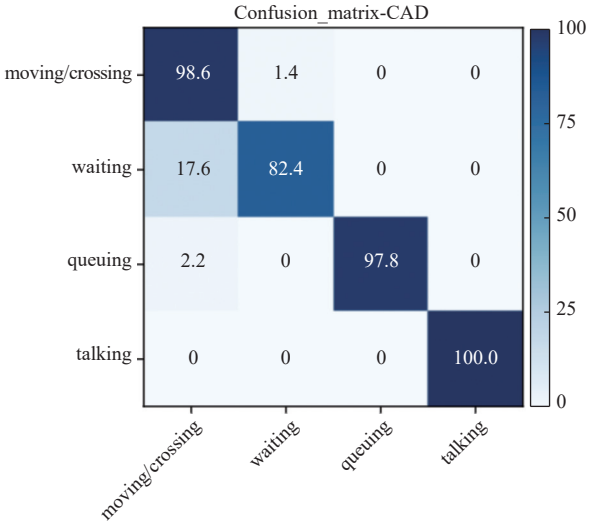


图 6　GIFFNet 在 Collective Activity 数据集上的混淆矩阵

图 7 展示了群体行为 talking 的可视化结果, 左图检测框中标注了目标人物, 右图展示了标注人物在时空层面上的动作权重, 颜色越深代表权重越大. 其中人物 0 在 10 个时间段内具有最大的平均权重, 是群体活动中的关键人物; 人物 3 是群体行为中的边缘人物, 具有最小的权重. 此外, 人物 0, 1, 2 具有明显的协作行为, 右图框出的权重颜色较深, 证明了他们在空间层面上的合作关系更为紧密. 在判别群体行为的过程中, 可能会出现目标群体包括多个行为类型的情况. 图 8 展示了群体行为 waiting 的可视化结果, 如左图所示, 标注的人物包括 waiting 与 walking 两种行为. 为了确定最终的群体行为类别, 分类器会根据各类行为的权重值大小, 选定权重最大的行为作为判别结果. 图 8 所示的示例经过分类器的选择, 判定为 waiting 的行为类别.

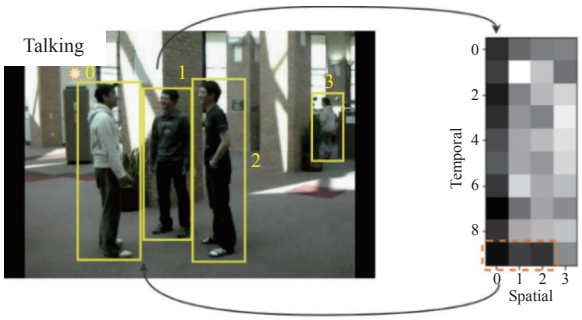


图 7　Collective Activity 数据集中行为 talking 的可视化示例

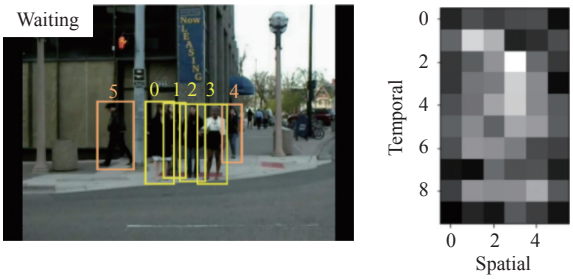


图 8　Collective Activity 数据集中 waiting 的可视化示例

4　结论与展望

针对群体行为识别时场景信息缺失的问题, 本文提出一种基于全局-个体特征融合的群体行为识别方法 GIFFNet. 总体来说, GIFFNet 分为两个阶段, 其一为特征提取阶段, 其二为特征融合与行为推理阶段. GIFFNet 中的全局-个体特征融合 (GIFF) 模块, 能够有效地提取全局特征中的场景信息, 并与个体特征进行融合, 达到增强特征表达能力的目的. 另外, GIFF 模块通过注意力机制以聚焦场景中的关键人物与数据集中的关键帧, 为群体行为的识别提供良好的准备条件. 为

确定融合时机的适配性, 本文研究了早期融合与晚期融合策略, 以总结出最佳的融合方式. 此外, 本文在标准交叉熵损失的基础上, 以动态分配权重的思想对损失函数做出了优化, 一定程度上解决了数据集样本失衡的问题. GIFFNet 在两个基线数据集 Volleyball 与 Collective Activity 上实现了高精度的群体活动识别, 其中多类分类准确度分别为 93.8%、96.1%, 类平均精确度分别为 93.9%、95.8%. 同时, GIFF 模块借鉴 bottleneck 模型的思想降低了计算复杂度, 但参数量依然较大. 在未来的研究中, 拟考虑实现更加轻量化的群体行为识别网络模型.

参考文献

1 吴建超, 王利民, 武港山. 视频群体行为识别综述. 软件学报, 2023, 34(2): 964–984. [doi: 10.13328/j.cnki.jos.006693]

2 Wu LF, Wang Q, Jian M, *et al*. A comprehensive review of group activity recognition in videos. International Journal of Automation and Computing, 2021, 18(3): 334–350. [doi: 10.1007/s11633-020-1258-8]

3 Ibrahim MS, Muralidharan S, Deng ZW, *et al*. A hierarchical deep temporal model for group activity recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1971–1980.

4 Tang JH, Shu XB, Yan R, *et al*. Coherence constrained graph LSTM for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(2): 636–647. [doi: 10.1109/TPAMI.2019.2928540]

5 Shu XB, Zhang LY, Sun YL, *et al*. Host-parasite: Graph LSTM-in-LSTM for group activity recognition. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(2): 663–674. [doi: 10.1109/TNNLS.2020.2978942]

6 Qi MS, Qin J, Li AN, *et al*. StagNet: An attentive semantic RNN for group activity recognition. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 104–120.

7 Shi L, Zhang YF, Cheng J, *et al*. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12018–12027.

8 Yuan HJ, Ni D, Wang M. Spatio-temporal dynamic inference network for group activity recognition. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 7456–7465.

9 Azar SM, Atigh MG, Nickabadi A, *et al*. Convolutional relational machine for group activity recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7884–7893.

10 Wu JC, Wang LM, Wang L, *et al*. Learning actor relation graphs for group activity recognition. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9956–9966.

11 Yuan HJ, Ni D. Learning visual context for group activity recognition. Proceedings of the 35th AAAI Conference on Artificial Intelligence. AAAI, 2021. 3261–3269.

12 Demirel B, Ozkan H. DECOMPL: Decompositional learning with attention pooling for group activity recognition from a single volleyball image. arXiv:2303.06439, 2023.

13 Gavrilyuk K, Sanford R, Javan M, *et al*. Actor-Transformers for group activity recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 836–845.

14 Pramono RRA, Chen YT, Fang WH. Empowering relational network by self-attention augmented conditional random fields for group activity recognition. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 71–90.

15 Cheng Y, Wang W, Ren ZP, *et al*. Multi-scale feature fusion and Transformer network for urban green space segmentation from high-resolution remote sensing images. International Journal of Applied Earth Observation and Geoinformation, 2023, 124: 103514. [doi: 10.1016/j.jag.2023.103514]

16 Choi W, Shahid K, Savarese S. What are they doing? Collective activity classification using spatio-temporal relationship among people. Proceedings of the 12th IEEE International Conference on Computer Vision Workshops. Kyoto: IEEE, 2009. 1282–1289.

17 Chappa NVS, Nguyen P, Nelson AH, *et al*. SPARTAN: Self-supervised spatiotemporal Transformers approach to group activity recognition. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Vancouver: IEEE, 2023. 5158–5168.

18 Yan R, Xie LX, Tang JH, *et al*. Social adaptive module for weakly-supervised group activity recognition. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 208–224.

19 Kim D, Lee J, Cho M, *et al*. Detector-free weakly supervised group activity recognition. Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition.

New Orleans: IEEE, 2022. 20051–20061.

20 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

21 Hu J, Shen L, Sun G. Squeeze-and-excitation networks. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 7132–7141.

22 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021.

23 Liu Z, Lin YT, Cao Y, *et al*. Swin Transformer: Hierarchical vision Transformer using shifted windows. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 9992–10002.

24 Mehta S, Rastegari M. MobileViT: Light-weight, general-purpose, and mobile-friendly vision Transformer. Proceedings of the Tenth International Conference on Learning Representations. OpenReview.net, 2022.

25 Mehta S, Rastegari M. Separable self-attention for mobile vision Transformers. arXiv:2206.02680, 2022.

26 Vasu PKA, Gabriel J, Zhu J, *et al*. FastViT: A fast hybrid vision Transformer using structural reparameterization. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 5762–5772.

27 Han MF, Zhang DJ, Wang YL, *et al*. Dual-AI: Dual-path actor interaction learning for group activity recognition. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 2980–2989.

28 Li SC, Cao QG, Liu LB, *et al*. GroupFormer: Group activity recognition with clustered spatial-temporal Transformer. Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021. 13648–13657.

29 Tamura M, Vishwakarma R, Vennelakanti R. Hunting group clues with Transformers for social group activity recognition. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 19–35.

30 Bagautdinov T, Alahi A, Fleuret F, *et al*. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 3425–3434.

31 Yan R, Tang JH, Shu XB, *et al*. Participation-contributed temporal dynamic model for group activity recognition. Proceedings of the 26th ACM International Conference on Multimedia. Seoul: ACM, 2018. 1292–1300.

32 Hu GY, Cui B, He Y, *et al*. Progressive relation learning for group activity recognition. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 977–986.

33 Wang DL, Liu J, Zhou Y. Group activity recognition based on temporal semantic sub-graph network. Proceedings of the 14th International Conference on Machine Learning and Computing. Guangzhou: ACM, 2022. 401–406.

34 Shu TM, Todorovic S, Zhu SC. CERN: Confidence-energy recurrent network for group activity recognition. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 4255–4263.

35 Yan R, Xie LX, Tang JH, *et al*. HiGCIN: Hierarchical graph-based cross inference network for group activity recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 6955–6968. [doi: 10.1109/TPAMI.2020.3034233]

(校对责编: 张重毅)