基于伪 3D 残差网络与交互关系建模的群组行为识别方法

丰 艳,张甜甜,王传旭
(青岛科技大学信息科学技术学院,山东青岛 266061)

摘 要: 针对复杂场景下群组行为特征的多样性以及交互关系难以建模的问题,提出一种全新的分层网络架构.第一层网络,利用伪 3D 残差网络与图卷积网络相结合捕获交互关系特征;第二层网络,利用伪 3D 残差网络捕获群组全局场景时空特征.根据上述特征之间的互补作用对它们的群组行为决策输出,提出一种权重自适应调整决策融合算法,对上面两层网络的群组行为类别自适应计算重要性权重,实现决策融合.该方法在 CAD 和 CAE 上分别取得了 91.4% 和 97.9% 的平均识别精度.

关键词: 群组行为识别;交互关系建模;自适应决策融合
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2020)07-1269-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.07.004

Group Activity Recognition Method Based on Pseudo 3D Residual Network and Interaction Modeling

FENG Yan, ZHANG Tian-tian, WANG Chuan-xu
(*School of Information Science & Technology, Qingdao University of Science and Technology, Qingdao, Shandong* 266061, *China*)

Abstract: For the diversity of group behavior characteristics in complex scenes and the problem of difficult interaction modeling, this paper proposes a new two-layered network architecture. The first layer of network combines a pseudo 3D residual network with a graph convolution network to capture the interaction characteristics. The second layer of network, uses the pseudo 3D residual network to capture the group global scene spatio-temporal characteristics. Based on the complementary role of the above features, their group behavior decisions are fused with a weight adaptive adjustment algorithm, which adaptively calculates importance weights for the group behavior categories predicted by the above two channels, and realizes decision fusion of the different prediction results. The method has achieved 91.4% and 97.9% average recognition accuracy on CAD and CAE respectively.

Key words: group behavior recognition; interaction modeling; adaptive decision fusion

1 引言

随着人工智能技术的不断发展,人体行为识别已经成为计算机视觉和模式识别领域的研究热点[1,2].它在体育视频分析、安防监控系统、虚拟现实以及人机交互等领域具有广阔的应用价值[3,4].目前,如何有效的描述人与人之间的交互关系,构建多个目标之间复杂的交互模型是一个具有挑战性的任务.

目前,基于深度学习的方法在群组行为识别中发挥了巨大作用.文献[5]提出一种结构化推断机制,利用 2D CNN 提取人体特征.由于 2D CNN 无法捕获连续帧间的时间信息,容易丢失关键的上下文信息.为了解决该问题,文献[6]提出一种 C3D 网络模型,虽然可以很好地捕获时空特征,但时间维度的引入,使整个网络的参数量和计算复杂度都大幅增加.此外,Karen 等人[7]提出一种基于双流网络的时空特征提取方法,在最终行为识别时采用加权平均方法,存在一定的机械性和盲目性.Deng 等人[8]提出了一种基于长短时记忆网络的分层模型网络,虽然能够有效提取人体时空特征,但忽略了人与人之间的交互关系特征.

收稿日期:2019-08-13;修回日期:2020-04-06;责任编辑:孙瑶
基金项目:国家自然科学基金面上项目(No.61672305);国家自然科学基金青年科学基金(No.61702295)

针对以上问题,本文做出以下两点贡献:①针对不同群组行为特征的多样性,提出一种分层模型,分别捕获交互关系特征与全局场景时空特征;②利用决策融合思想,根据不同特征在群组行为识别中的互补作用,提出一种权重自适应调整决策融合算法,对基于不同特征得到的行为类别自主赋予权值,实现复杂场景下的群组行为识别.

2 相关工作

针对现有研究方法中是否考虑目标交互关系建模,可以分为两大类:无图模型群组行为识别方法和基于图模型的群组行为识别方法.

第一大类无图模型方法研究比较早,通常将整个场景作为一个整体进行群组行为分析. Vahora[9] 等人提出一种基于关系内容的学习模型,采用自下而上的方法从活动场景中依次学习单人行为和群组行为. Ramanathan 等人[10] 提出一种基于注意力机制的群组行为识别方法,使用循环神经网络捕获关键人物的时序信息对群组行为进行识别. 第二大类基于图模型交互关系建模的方法,聚焦于群组内成员之间的交互关系建模,以期提高群体行为的识别精度. Li 等人[11] 提出一种超图内聚类搜索算法. Deng 等人[12] 提出一种基于概率图模型的深度神经网络结构. Wu 等人[13] 利用图卷积网络(Graph Convolutional Networks, GCN)捕获交互关系特征,最后利用特征融合[14] 的方法进行群组行为识别. 此外,在进行群组行为识别时,与外观、光流和轨迹[15] 模态相比,人体的骨骼姿态图较少受到关注,但在进行人体目标检测时,骨骼姿态图比边界框定位更加准确.

受以上方法启发,本文基于人体骨骼姿态图进行时空特征提取,利用 GCN 捕获人与人之间的交互关系特征,并对该特征进行动态维护.

3 整体算法框架概述

图 1 为本文提出算法的整体框架示意图,将群组行为识别过程划分为四个阶段.

Stage1 视频预处理. 采用 OpenPose 姿态估计算法[16] 对原始视频进行预处理,经过不同处理机制得到两种不同类型的姿态图,一种是单人姿态图,另一种是场景姿态图.

Stage2 交互关系特征提取. 将预处理得到的单人姿态图输入伪 3D 残差网络[17](Pseudo 3D Residual network, P3D ResNet),提取单人时空特征,然后利用单人时空特征以及位置信息构建交互关系无向图,再利用 GCN 网络对交互关系进行动态维护,捕获人与人之间的交互关系特征.

Stage3 全局场景时空特征提取. 将预处理得到的场景姿态图输入 P3D ResNet 网络中,提取全局场景时空特征.

Stage4 决策融合. 将交互关系特征和全局场景时空特征分别输入两个 Softmax 分类器中,获得两个预测结果. 最后,采用权重自适应调整决策融合算法对上述两种预测结果进行决策融合并输出最终群组行为类别.

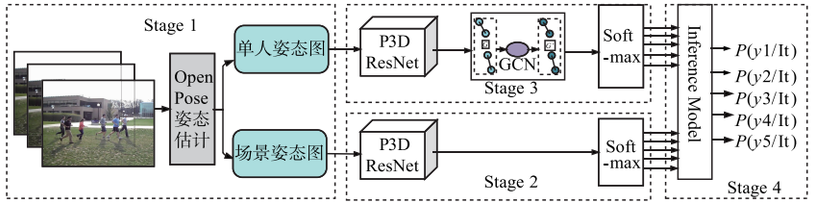


图1 群组行为识别算法示意图

4 场景时空特征提取

4.1 伪 3D 残差网络

P3DResNet 将 3D 卷积中 3 * 3 * 3 的时空卷积核分解为 1 * 3 * 3 的空间卷积核和 3 * 1 * 1 的时间卷积核. 这种方法极大的降低了网络参数量,从而可以提升网络运行速度. 本文所采用的 P3D ResNet199 是基于 ResNet 152 得到的,网络深度之所以增加是因为残差结构不再是 3 个卷积层而是 4 个,如图 2 所示.

P3DResNet 网络由大量残差单元构成,残差单元用以下方式来进行表示:

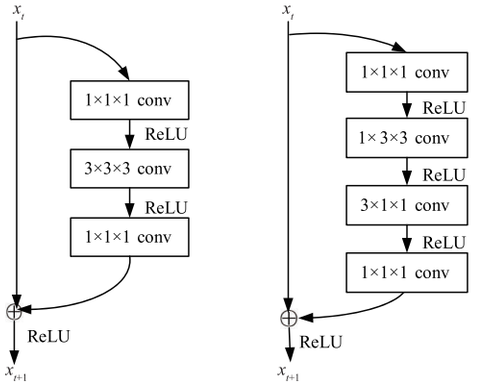
$$x_{m+1} = h(x_m) + F(x_m) \tag{1}$$

$$x_M = x_1 + \sum_{m=1}^{M-1} F(x_m) \tag{2}$$

其中,x_m 表示第 m 个残差单元的输入,x_{m+1} 表示第 m 个残差单元的输出,$h(x_m) = x_m$ 表示恒等映射关系,F 是非线性残差函数,由图 2(b)中右侧包含的四个卷积操作构成. 本文设计采用 M 个时空特征提取模块 $M = 33$,如式(2)所示,对两种预处理的图像进行时空特征提取. 每一部分特征提取的详细过程将在 4.2 和 5.1 小节进行描述.

4.2 场景时空特征提取

利用 P3D ResNet 网络,输入一组连续的场景姿态图 $l = \{l_1, l_2, \cdots, l_T\}$,$T$ 为视频帧数,设定 $T = 16$. 首先,



(*a*) 3D ResNet 152 残差块　　(*b*) P3D ResNet 199 残差块

图2　ResNet 152与P3D ResNet 199残差结构对

将输入层或时空特征提取层的输出经过一个 1 * 1 * 1 的卷积层调整特征维度;然后输入 1 * 3 * 3 的空间卷积层提取空间特征,进一步输入 3 * 1 * 1 的时间卷积层得到时空特征;最后,使用 1 * 1 * 1 的卷积核进行特征融合得到场景时空特征表示. 在经过 28 个时空特征提取模块后,通过平均池化层和全连接层得到每个视频帧的场景特征描述符 $S=(s_1,s_2,\cdots,s_T)$,然后输入 Softmax 分类器中计算出每个视频帧在所有群组活动类别上的概率预测值 $L=\{l_1,l_2,\cdots,l_Q\}$,Q 表示群组行为的类别总数.

5　群组交互关系建模

5.1　单人时空特征提取

利用 P3D ResNet 网络,输入一组连续的场景姿态图 $\{P_1,P_2,\cdots,P_N\}$,N 为视频中的人体总数. 与 4.2 节场景时空特征提取过程相同,在经过 33 个时空特征提取模块后,通过平均池化层和全连接层为每个人产生一个时空特征向量 $\boldsymbol{X}^a=(x_1^a,x_2^a,\cdots,x_N^a)$,进而通过 Softmax 分类器预测每个人的行为类别 $y=\{y_1,y_2,\cdots,y_N\}$.

5.2　构建群组关系无向图模型

利用 5.1 节每个人的时空特征以及位置特征构建交互关系无向图 $G=(V,E)$,如图 3 所示,图中顶点 Object_i 表示单个人的特征 $V=\{(x_i^a,x_i^p)\mid i=1,2,\cdots,N\}$,$x_i^a\in\mathbf{R}^d$ 表示第 i 个人的时空特征,由 P3D ResNet 对相邻四帧图像提取特征得到,不同颜色的顶点代表不同行为类别. $x_i^p=(t_i^x,t_i^y)$ 表示第 i 个人的中心位置坐标,由 OpenpPose 算法输出的骨骼中心节点的平均位置坐标来表示. E 是连接两个顶点的边.

从图 3(*a*)可以看出,由于 Object_1、Object_2 和 Object_3 之间的位置距离较近,因此在单人行为识别过程中误将 Object_1 识别为“Talking”. 在经过 GCN 对节点特征重提取后,对判别错误的行为进行了纠正,用图 3(*b*)中的绿色节点表示,其正确行为类别为“Walking”;从图 3(*a*)到图 3(*b*)再到图 3(*c*),每经过一次 GCN 处理,相邻节点特征就进行一次融合. 在此过程中,节点颜色逐渐加深表示对该节点所属行为的置信度逐渐增加.

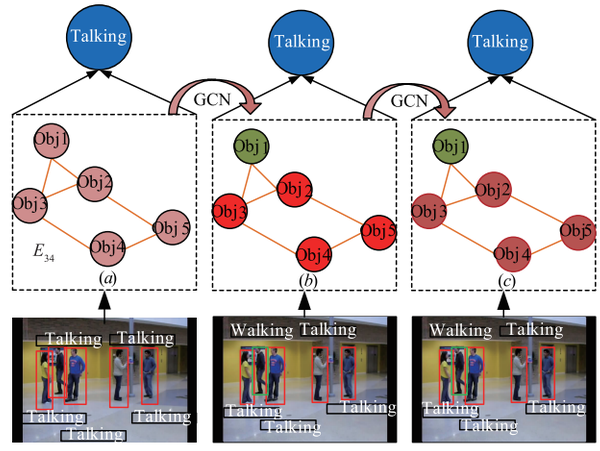


图3　群组关系无向图

5.3　交互关系度量

交互关系度量分为时空特征度量和位置关系度量.

时空特征度量. 本文采用向量点积的方法计算两个体之间的时空特征关系,其计算公式如下:

$$f_a(x_i^a,x_j^a)=\frac{(x_i^a)^T x_j^a}{\sqrt{d}} \tag{3}$$

其中$\sqrt{d}$为归一化因子.

位置关系度量. 本文采用欧式距离公式计算个体之间的位置关系,计算公式如下:

$$f_p(x_i^P,x_j^P)=\text{II}(d(x_i^P,x_j^P)\leqslant\mu) \tag{4}$$

其中 II(·)是一个指标函数;$d(x_i^s,x_j^s)$ 表示两个人边界框中心点之间的欧式距离;μ 为阈值,取值为图像宽度的 1/5. 当两个人之间的位置距离小于 μ 时,说明两个人之间的关系较弱,去掉节点之间的连线.

根据人体目标的特征,计算任意两个节点之间边的权重,如图 3 所示,距离较近且时空特征相似性高的两个节点关系较强,用粗线表示;相反,用细线表示. 该交互关系权重用函数 $E_{ij}\in\mathbf{R}^{N\times N}$ 表示如下:

$$E_{ij}=F(f_a(x_i^a,x_j^a),f_p(x_i^p,x_j^p)) \tag{5}$$

其中,$f_a(x_i^a,x_j^a)$ 表示人体目标 i 和 j 之间的时空特征关系,$f_p(x_i^p,x_j^p)$ 表示其位置关系,F 是一个复合函数,作用是将时空特征和位置关系融合成一个标量权重,其具体的计算公式如下:

$$E_{ij}=\frac{f_p(x_i^p,x_j^p)\exp(f_a(x_i^a,x_j^a))}{\sum_{j=1}^{N}f_p(x_i^p,x_j^p)\exp(f_a(x_i^a,x_j^a))} \tag{6}$$

采用 Softmax 函数对每个节点的交互权重进行归

一化,得到的交互关系特征为使每个节点的对外交互关系权重的总和为 1.

5.4 交互关系动态维护

本算法采用 GCN 对交互关系进行动态维护,具体流程如图 4 所示. 在动态维护过程中可堆叠多个 GCN 块,形式上一层 GCN 结构可以用如下公式来表示:

$$Z^{(l+1)}=\sigma(\boldsymbol{G}Z^{(l)}\boldsymbol{W}^{(l)}) \tag{7}$$

其中,$\boldsymbol{G}\in\mathbf{R}^{N\times N}$是图的矩阵表示;$Z^{(l)}\in\mathbf{R}^{N\times d}$是第 l 层节点的特征表示,且 $Z^{(o)}=\boldsymbol{A}$,$\boldsymbol{A}$ 为 $\boldsymbol{G}$ 的邻接矩阵;$\boldsymbol{W}^{(l)}\in\mathbf{R}^{d\times d}$是第 l 层可学习到的权重矩阵;$\sigma(\cdot)$是 ReLU 函数.

本文模型共堆叠了七个 GCN 块,每个 GCN 块中都包含图卷积层、归一化层和 ReLU 层. 在实验过程中,根据对每层内核大小的研究,将前两层的卷积核尺寸设计为 $3*1$;第三层设计为初始结构,使用 $1*1$ 的卷积核;第四层和第六层仍然使用 $3*1$ 的卷积核;第五层和第七层使用 $1*1$ 的卷积核,经过全连接层得到的交互关系特征向量为 $\boldsymbol{G}'=\{g_1,g_2,\cdots,g_N\}$,进而将交互关系特征向量输入 Softmax 分类器中计算基于交互关系的群组活动预测值.

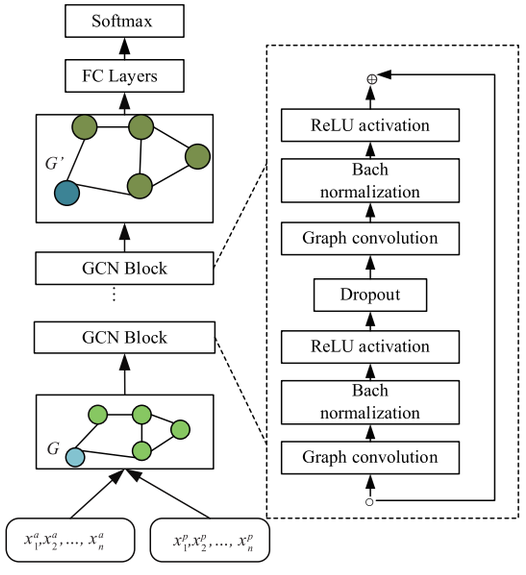


图4 利用GCN进行交互关系动态维护

6 权重自适应调整决策融合

本文根据交互关系特征与全局场景特征在不同群组行为识别中所起的互补作用,提出一种权重自适应调整的决策融合方法,通过对不同行为类别自主赋予权重,并对不同分类器的结果进行融合,输出最终分类结果. 具体迭代过程如下.

步骤 1 初始化每个行为类别 $C\in\{c_1,c_2,\cdots,c_n\}$下不同分类器 $U\in\{u_1,u_2,\cdots,u_m\}$输出的权重 $W\in\{W^{c_1},W^{c_2},\cdots,W^{c_n}\}$,$W^{c_i}\in\{W^{c_i}_{u_1},W^{c_i}_{u_2},\cdots,W^{c_i}_{u_m}\}$,计算公式为

$$W^{c_i}_{u_j}=1/z \tag{8}$$

1 其中,n 为行为类别数量,z 为分类器数量,c_i 表示第 i 个类别,u_j 表示第 j 个分类器. 本文 $z=2$,初始化交互关系权重系数 $W^{c_i}_{u_j}=0.5$.

步骤 2 对于任意一个视频样本 $x\in X$,分别计算每个分类器下不同分类器的预测概率 $P\in\{P^{c_1},P^{c_2},\cdots,P^{c_n}\}$,$P^{c_i}\in\{P^{c_i}_{u_1},P^{c_i}_{u_2},\cdots,P^{c_i}_{u_m}\}$其中 $P^{c_i}_{u_j}$表示第 j 个分类器预测 x 是第 i 个行为类别的概率.

步骤 3 利用权重 W 对每个行为类别下不同分类器的预测概率值进行加权融合,计算每个行为类别的累加概率:

$$P^{c_i}=\sum_{j=1}^{m}w^{c_i}_{u_j}P^{c_i}_{u_j} \tag{9}$$

$$\Gamma=\{P^{c_i}\mid i=1,2,\cdots,n\} \tag{10}$$

步骤 4 通过决策融合后的最大概率,确定输入视频片段所属的群组行为类别:

$$P(x)=\arg\max_i\{\Gamma\};i=1,2,\cdots,n \tag{11}$$

步骤 5 自适应迭代更新 $w^{c_i}_{u_j}$,具体方法为:假设样本 x 的真实标签为 $L(x)$,决策融合后的分类标签为 $Y(x)$,若 $L(x)=Y(x)$,将分类器输出的概率预测值 $P^{c_i}_{u_j}$从大到小倒序排列,选择预测错误的 d 个分类器,将其对应的权重系数减去 ε,同时将前 d 个分类器所对应的权重系数加上 ε,得到更新后的权值 $w^{*c_i}_{u_j}$,本文设定 $\varepsilon=0.05$. 若 $L(x)\neq Y(x)$,则判断该样本为噪声,直接丢弃,并返回步骤 2,直到遍历完成所有样本. 经过以上 5 个步骤,整个决策融合的模型训练完毕,并将最大后验概率值所对应的群组行为类别作为最终的群组行为类别.

7 实验与评估

7.1 数据集

CAD 数据集[18]包含 44 个视频片段,由低分辨率手持相机拍摄. 共包含 5 种群组活动标签:Crossing、Queuing、Walking、Talking 和 Waiting. 所有视频序列,每 10 帧标注一次,标注的信息包括人体的边界框和行为标签. 数据集中每种行为被随机分为训练集和测试集. 其中 75% 的样本用于训练,15% 用来测试. CAE 数据集[19]是 CAD 的扩展集,共有 33 个视频片段. 在删掉 Walking 这个模糊的行为后,又引入了 Jogging 和 Dancing 两种新行为. 数据集的划分与 CAD 相同.

7.2 实验配置及网络参数设置

本实验在硬件上使用一个 Nvidia 1080Ti GPU 对模型进行训练. 由于内存的限制,将模型分为两个阶段进行训练. 整个模型使用 PyTorch 框架实现. 采用 ADAM 算法学习网络参数. 对于 CAD,在网络训练时采用的最小批量为 32 帧,共迭代 100 次,初始学习率设置为

0.01,每迭代 10 次降低 0.1 倍. 对于 CAE,采用的最小批块为 16 帧,共迭代 35 次,初始学习率设置为 0.001,每迭代 10 次降低 0.1 倍.

7.3　基线模型对比

7.3.1　Baselne 模型设计

为了验证本文整体网络架构中各个部分的性能,共设计四种基线模型与本文模型进行比较.

Baseline1 为 RGB 图 + 交互关系特征;

Baseline2 为 RGB 图 + 全局场景时空特征;

Baseline3 为 RGB 图 + 交互关系特征 + 全局场景时空特征 + 加权平均融合;

Baseline4 为 RGB 图 + 交互关系特征 + 全局场景时空特征 + 权重自适应调整决策融合;

本文算法为姿态图 + 交互关系特征 + 全局场景时空特征 + 权重自适应调整决策融合.

7.3.2　实验结果分析

图 5 显示的是四种基线模型以及本文模型在 CAD 上不同行为识别精度对比结果以及最终的平均识别精度(MPCA).

经过实验结果对比,可以得出以下几点结论:

(1)通过比较 Baseline1 与 Baseline2 可以说明交互关系特征比全局场景时空特征的性能更优.

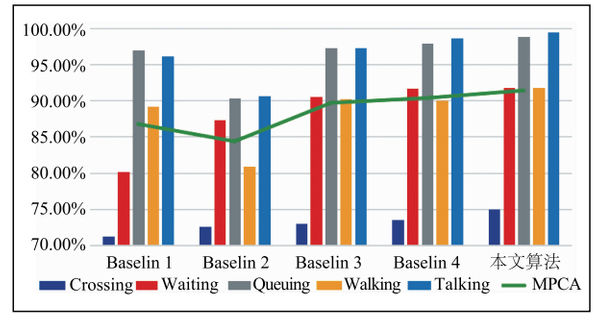


图5　本文算法与四种基线模型在CAD数据集上的实验对比结果

(2)将 Baseline3 与 Baseline1 和 Baseline2 对比说明特征加权融合后的结果高于各自独立的结果.

(3)将 Baseline4 与 Baseline3 的实验结果对比说明权重自适应调整决策融合算法性能优于加权平均融合算法.

(4)将本文算法与 Baseline4 的实验结果对比,证明了本文算法优于四种基线模型.

7.4　本文模型与当前流行方法在 CAD 数据集上的对比实验

表 1 显示了本文算法与四种流行方法在 CAD 数据集上的识别精度对比结果. 包括每种方法在不同行为类别上的识别精度以及平均识别精度.

表 1　本文方法与其他流行方法在 CAD 数据集上的识别率对比

Model	Crossing	Waiting	Queuing	Walking	Talking	MPCA
Hierarchical Model(RGB)[8]	61.54	66.44	96.77	80.41	99.45	81.5%
SBGAR(RGB + 光流)[20]	78.03	81.37	99.16	87.58	84.62	86.1%
FTC-HiRF[21]	86.5	85.9	98.2	89.7	99.6	**92.0%**
CNN + GCN(RGB)[13]	–	–	–	–	–	91.0%
CNN + GCN(RGB + OpenPose)	74.05	91.82	98.89	91.8	99.44	91.2%
P3DResNet + GCN(RGB)	73.72	91.81	98.8	91.78	99.39	91.1%
P3DResNet + GCN(Ours)	74.99	91.84	98.9	91.82	99.45	91.4%

通过对比表 1 所有方法的平均识别精度可以得出以下几点结论:

(1)通过与文献[8,20]实验结果对比,说明姿态图比 RGB 图包含的信息量更丰富. 考虑交互关系建模比无交互关系建模的识别率高.

(2)与 FTC-HiRF[21] 实验结果相比,说明自底向上与自顶向下的联合推断方式可以提高群组行为识别精度.

(3)与 CNN + GCN[13] 算法相比,说明分层网络模型性能优于单层网络模型.

(4)利用 OpenPose 算法在 Bounding box 中叠加骨架信息,比没有骨架信息的识别率更高.

(5)本文利用多路特征互补的原则,通过权重自适应调整决策融合算法实现预测结果的决策融合,可以有效提高群组行为识别精度.

通过对比不同行为类别上的识别精度可以发现,本文算法对“Queuing”和“Talking”两种行为的识别精度接近 100%,而对“Crossing”的识别精度较低. 主要原因是“Crossing”和“Walking”之间存在极大的相似性,且二者经常发生在同一个场景中. 因此,为了提高识别的准确性,在 CAE 中将“Crossing”和“Walking”两种行为进行合并,统称为“Moving”.

7.5　本文算法与当前流行方法在 CAE 数据集上的对比实验

表 2 比较了本文算法和当前已有算法在 CAE 数据集上的实验结果. 相比而言,本文算法获得了更为优越

的识别效果.

表 2 本文与当前流行方法在 CAE 数据集上的识别率对比

Method	Accuracy(%)
UTR [22]	80.8
CRF + CNN [23]	86.7
Hypergraphs Model [11]	95.1
Structure Inference Machines [5]	90.2
P3D ResNet + GCN(Ours)	97.9

本文算法与四种流行算法都是基于交互关系的群组行为识别方法,通过对比可以得出以下几点结论:

(1)文献[22]采用人工设计方法提取的特征,受限于设计者的先验知识,具有很大的局限性,因此,与本文 EC3D 的方法相比其识别率相对较低.

(2)文献[23]采用条件随机场提取交互关系特征,由于条件随机场只能编码视频特征中的短期依赖关系,因此与本文所采用的可捕获长期依赖关系的 GCN 相比,识别精度较低.

(3)文献[11]在进行特征提取之前对数据没有进行预处理,而本文通过 OpenPose 的方法提取人体姿态信息,缩小了感兴趣目标,使提取的特征更加精细,因此识别精度高.

(4)本文单独进行了全局场景时空特征提取,丰富了语义信息.

(5)与文献[5]采用单一特征进行群组行为识别的方法相比,本文采用多线索特征综合利用的方式可以有效提高群组行为识别精度.

图 6 展示了本文算法分别在 CAD 和 CAE 两个数据集上的识别效果图,从图中可以看出,识别的内容包含每个人的行为类别以及群组行为类别. 编号 1 ~ 7 代表单人行为标签,每张图片左上角的标签代表该视频帧所属的群组行为类别.

图6 本文算法在CAD和CAE两个数据集上的识别效果图

8 结论

本文针对复杂场景下群组行为特征的多样性以及交互关系难以建模的问题,提出一种全新的分层网络架构. 利用 P3D ResNet 与 GCN 相结合分别提取交互关系特征和全局场景时空特征,并描述了如何进行交互关系建模以及交互关系的动态维护. 根据不同特征之间的互补作用,提出一种权重自适应调整决策融合算法,在 CAD 和 CAE 两个数据集上的实验结果同时表明了本文算法的有效性. 未来计划通过引入注意力机制实现对关键人物的识别,并将其扩展到更大的数据集中,解决更为复杂的群组行为识别问题.

参考文献

[1] 韩磊,李君峰,贾云得. 基于时空单词的两人交互行为识别方法[J]. 计算机学报,2010,33(4):776 – 784.
HAN Lei, LI Jun-feng, JIA Yun-de. Human interaction recognition using spatio-temporal words[J]. Chinese Journal of Computers, 2010, 33(4): 776 – 784. (in Chinese)

[2] 朱煜,赵江坤,王逸宁,郑兵兵. 基于深度学习的人体行为识别算法综述[J]. 自动化学报,2016,42(6):848 – 857.
ZHU Yu, ZHAO Jiang-kun, WANG Yi-ning, ZHENG Bing-bing. A review of human action recognition based on deep learning[J]. Acta Automatica Sinica, 2016, 42(6): 848 – 857. (in Chinese)

[3] 郑兴华,孙喜庆,吕嘉欣,等. 基于深度学习和智能规划的行为识别[J]. 电子学报,2019,47(8):1661 – 1668.
ZHENG Xing-hua, SUN Xi-qing, LU Jia-xin, et al. Action recognition based on deep learning and artificial intelligence planning[J]. Acta Electronica Sinica, 2019, 47(8): 1661 – 1668. (in Chinese)

[4] 王传旭,刘云,厉万庆. 基于时空特征点的非监督姿态建模和行为识别的算法研究[J]. 电子学报,2011,39(8):1751 – 1756.
WANG Chuan-xu, LIU Yun, LI Wan-qing. Research ofunsupervised posture modeling and action recognition based on spatial-temporal interesting points[J]. Acta Electronica Sinica, 2011, 39(8): 1751 – 1756. (in Chinese)

[5] Deng Z, Vahdat A, Hu H, et al. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 4772 – 4781.

[6] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. USA: IEEE, 2015. 4489 – 4497.

[7] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[A]. Advances in Neural Information Processing Systems[C]. USA: Massachusetts Institute of Technology Press, 2014. 568 – 576.

[8] Ibrahim M S,Muralidharan S,Deng Z,et al. A hierarchical deep temporal model for group activity recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 1971 – 1980.

[9] Vahora S, Chauhan N. Deep neural network model for group activity recognition using contextual relationship [J]. Engineering Science and Technology,an International Journal,2019,22(1):47 – 54.

[10] Ramanathan V,Huang J,Abu-El-Haija S,et al. Detecting events and key actors in multi-person videos[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2016. 3043 – 3053.

[11] Li W,Chang M-C,Lyu S. Who did what at where and when:simultaneous multi-person tracking and activity recognition[J]. arXiv Preprint,2018,arXiv:1807. 01253.

[12] Deng Z,Zhai M,Chen L,et al. Deep structured models for group activity recognition[J]. arXiv Preprint,2015,arXiv:1506. 04191.

[13] Wu J, Wang L, Wang L, et al. Learning actor relation graphs for group activity recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2019. 9964 – 9974.

[14] 罗会兰,王婵娟. 行为识别中一种基于融合特征的改进 VLAD 编码方法[J]. 电子学报,2019,47(1):49 – 58.
LUO Hui-lan, WANG Chan-juan. An improved VLAD coding method based on fusion feature in action recognition[J]. Acta Electronica Sinica,2019,47(1):49 – 58. (in Chinese)

[15] 田国会,尹建芹,闫云章,李国栋. 基于混合高斯模型和主成分分析的轨迹分析行为识别方法[J]. 电子学报,2016,44(1):143 – 149.
TIAN Guo-hui, YIN Jian-qin, YAN Yun-zhang, LI Guodong. Gaussian mixture models and principal component analysis based human trajectory behavior recognition[J]. Acta Electronica Sinica,2016,44(1):143 – 149. (in Chinese)

[16] Cao Z,Simon T,Wei S-E,et al. Realtime multi-person 2d pose estimation using part affinity fields[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA:IEEE,2017. 7291 – 7299.

[17] Qiu Z,Yao T,Mei T. Learning spatio-temporal representation with pseudo-3d residual networks[A]. Proceedings of the IEEE International Conference on Computer Vision [C]. USA:IEEE,2017. 5533 – 5541.

[18] Choi W, Shahid K, Savarese S. What are they doing?: Collective activity classification using spatio-temporal relationship among people[A]. IEEE 12th International Conference on Computer Vision (ICCV) Workshops [C]. USA:IEEE,2009. 1282 – 1289.

[19] Choi W,Shahid K,Savarese S. Learning context for collective activity recognition[A]. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern RecognitionJune(CVPR'11)[C]. USA: IEEE, 2011. 3273 – 3280.

[20] Li X,Choo Chuah M. SBGAR:semantics based group activity recognition[A]. Proceedings of the IEEE International Conference on Computer Vision[C]. USA:IEEE, 2017. 2876 – 2885.

[21] Lan T, Wang Y, Yang W, et al. Discriminative latent models for recognizing contextual group activities[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2011,34(8):1549 – 1562.

[22] Choi W,Savarese S. Understanding collective activitiesof people from videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(6): 1242 – 1257.

[23] Amer M R,Lei P,Todorovic S. Hirf:Hierarchical random field for collective activity recognition in videos[A]. European Conference on Computer Vision[C]. Cham: Springer,2014. 572 – 585.

作者简介

丰　艳　女,1977 年 10 月出生,山东曲阜人. 现为青岛科技大学副教授,硕士生导师. 主要从事虚拟现实、计算机视觉方面的研究.
E-mail:fywmh@163.com

张甜甜　女,1993 年 3 月出生,山东烟台人. 2017 年毕业于齐鲁工业大学信息学院,取得计算机科学与技术专业学士学位,现为青岛科技大学信息学院在读硕士研究生,从事计算机视觉方面的有关研究.
E-mail:zhangtt0424@163.com

王传旭　男,1968 年 1 月出生,山东邹城人. 教授、硕士生导师. 1990 年、2000 年和 2007 年分别在中国石油大学(华东)、中国石油大学(北京)和中国海洋大学获应用电子技术学士、硕士和博士学位. 主要从事计算机视觉方面的有关研究.
E-mail:Wangchuanxu_qd@163.com