基于 GFU 和分层 LSTM 的组群行为识别研究方法

王传旭,薛 豪
(青岛科技大学信息科学技术学院,山东青岛 266001)

摘 要: 提出一种以“关键人物”为核心,使用门控融合单元(GFU,Gated Fusion Unit)进行特征融合的组群行为识别框架,旨在解决两个问题:①组群行为信息冗余,重点关注关键人物行为特征,忽略无关人员对组群行为的影响;②组群内部交互行为复杂,使用 GFU 有效融合以关键人物为核心的交互特征,再通过 LSTM 时序建模成为表征能力更强的组群特征. 最终,通过 softmax 分类器进行组群行为类别分类. 该算法在排球数据集上取得了 86.7% 的平均识别率.

关键词: 组群行为识别; 关键人物建模; 交互特征建模; 门控融合单元
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112 (2020)08-1465-07
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.08.002

Group Activity Recognition Based on GFU and Hierarchical LSTM

WANG Chuan-xu, XUE Hao
(*Institute of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong* 266001, *China*)

Abstract: This paper proposes a group behavior recognition framework with “key persons” as the core and Gated Fusing Unit (GFU) for feature fusion. Its aim is to solve the following two problems :1) Group behavior information is redundant, focusing on key person behavior characteristics, ignoring the influence of unrelated personels on group behavior. 2) The internal interaction relationship is complex within group, GFU is used to effectively model interaction feature centered around the key characters and it is temporally evolved into the group characteristics via LSTM processing. Finally, the group behavior category is classified with Softmax. The algorithm achieves an average recognition rate of 86.7% on the volleyball dataset.

Key words: group behavior recognition; key person modeling; interaction feature modeling; gated fusion unit

1 引言

目前组群行为识别方法大概分为两个步骤:①提取个人特征分析个人动作;②将个人动作进行聚合推断出组群行为标签. 在第①步中,文献[1]中提出使用卷积神经网络对个人特征进行提取,以此作为 LSTM 网络地输入,以更好地理解个体的动作. 在第②步中,文献[2]通过图模型和递归神经网络(RNN)对场景中的人与人之间的高阶关系进行编码以获得组群关系特征实现分类. 但是,以上这些方法都是等同地利用每个成员动态特征,将整个组群看成一个整体进行组群行为识别的,忽略了组群成员个体对组群行为识别贡献差异这个问题.

Deng 等人[3]提出了一种将图模型和深度神经网络集成到联合框架中的方法,验证了组群中只有少数参与者在整个组群活动中发挥主导作用,因此,本文以这些关键人物为核心,提出了一种基于门控融合单元 GFU(Gated Fusion Unit)和长短时记忆网络 LSTM(Long Short-term Memory Network)的网络框架,主要贡献如下:

(1)提出了一种组群行为中的“关键人物”建模方法,堆叠每个人的光流图像,计算平均运动强度,按降值排序确定关键人物,在本文中将其称为具有长时间稳定运动的个体.

(2)提出了一种以“关键人物”为核心的组群特征

收稿日期:2019-09-19;修回日期:2020-03-31;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61672305)

融合方法,使用门控融合单元融合个人特征和场景特征,捕捉与关键人物相关的子群体的交互特征,更高效地完成组群行为识别.

2 相关工作

组群行为是由多人协同完成的,因此,以个人特征和交互特征推断组群行为是研究重点. Ibrahim 等人[1]提出了一种两阶段 LSTM 的分层模型. 第一层 LSTM 网络对个人层级的特征进行识别;第二层 LSTM 网络使用最大池化策略对个人特征进行取舍融合,旨在以点带面表征群体特征进行组群行为识别. Shu 等人[4]提出置信能量递归网络(CERN),将个体行为预测和交互行为预测的置信度与能量层相结合,推断出组群的类别标签. 文献[5]提出一种结构化递归神经网络(SRNN),使用一系列相连的 RNN 共同捕获个人行为、交互特征以及群体活动,以此推断场景中的哪些人正在进行交互并推断出组群行为的标签. Denman 等人[6]将个人特征融合生成"动作代码"表示组群特征,该模型基于 GAN 网络自动学习损失函数的能力完成组群行为识别. 此外,部分研究人员[7~9]使用图模型进行组群行为识别也取得了不错的结果.

上述方法的特点都是着重对组群成员个体特征进行建模,通过描述交互关系来表达组群整体特征,但是未对组群成员的贡献大小进行区分对待,进而影响整个组群特征的描述精度,影响识别效果. 针对这一缺点,人们使用注意力模型关注组群当中贡献更大的人员,Yan 等人[10]使用分层 LSTM 网络架构,并引入注意力机制关注对组群行为识别做出突出贡献的个体. Kong 等人[11]在个人层面应用了"分层注意网络"可以区分不同的人及其身体部分,以识别出关键人物. 文献[12]提出了一个基于注意力机制模型,能够自动地关注到和事件发生最相关的那部分人. 研究表明,对组群成员贡献进行区分,极大的提高了组群行为识别的精度.

3 算法描述

3.1 总体网络架构概述

关键人物特征以及交互特征在组群行为识别中起到了重要的作用,因此,本文以这些特征为核心进行组群行为识别,图 1 为算法流程图,概述如下.

首先,对场景中的个体进行跟踪,得到个体的边界框,使用预训练的 CNN 网络提取个体以及相对应场景图像的静态(RGB)特征. 其次,对个体和场景静态特征使用 LSTM 网络进行动态特征提取,其中,本文使用光流特征计算个体的平均运动强度,规定运动强度大的个体为关键人物. 然后,使用门控融合单元(Gate Fusion Unit,GFU)对关键人物以及个体间的交互信息进行融合. 根据输入的场景图像,可以确定个体在场景中的空间位置信息,用以表示组群内部的交互信息. 融合过程中,对组群行为作出突出贡献的个体会获得更大的权重. 最后,将融合后的特征输入到组群级 LSTM,捕捉组群级别的动态,并且连接 softmax 分类器进行分类.

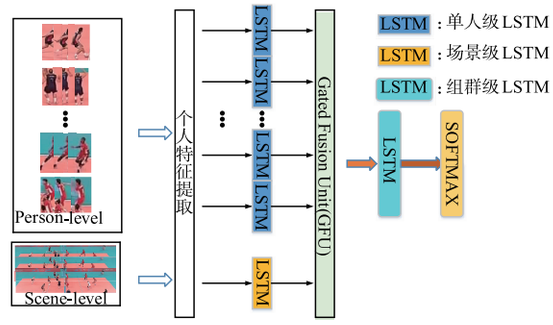

图1 基于GFU和分层LSTM的组群行为识别流程图

3.2 特征提取

3.2.1 静态特征提取

文献[13]中提出一种基于 CNN 网络的多信息流动卷积神经网络模型提取行人特征,本文使用 DSST 跟踪算法[14]得到个人边界框 $X=\{x_1^n,x_2^n,\cdots,x_t^n\}$,$n\in[1,N]$,$N$ 代表场景中的总人数,当 $n=1$ 时,代表场景中第一个人的特征序列. 跟踪到的场景级输入为 $\hat{X}=\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_t\}$,使用在 Image-Net[15]上预训练的 ResNet-50[16]网络模型进行静态特征提取:

$$I_G=(\{x_1^1,x_2^1,\cdots,x_T^1\},\cdots,\{x_1^N,x_2^N,\cdots,x_T^N\},\{\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_T\}) \tag{1}$$

提取的场景级静态特征为:

$$\hat{\theta}_t=f(\hat{X}_t) \tag{2}$$

提取的个人静态特征为:

$$\theta_t^n=f(x_t^n) \tag{3}$$

将个人静态特征 θ_t^n 和场景级静态特征 $\hat{\theta}_t$ 输入到第一层 LSTM 网络进行动态特征提取.

3.2.2 动态特征提取

文献[17]采用 LSTM 网络来充分利用上下文信息,提升模型效果. 本文同样采用 LSTM 网络来学习场景级和个人级的动态特征,其输入门 i_t,遗忘门 f_t,输出门 o_t 和输入调制门 g_t 以及单人 LSTM 的存储单元 c_t 的定义如下:

$$i_t=\sigma(\boldsymbol{W}_{ix}[h_{t-1},\theta_t]+\boldsymbol{b}_i) \tag{4}$$

$$f_t=\sigma(\boldsymbol{W}_{fx}[h_{t-1},\theta_t]+\boldsymbol{b}_f) \tag{5}$$

$$o_t=\sigma(\boldsymbol{W}_{ox}[h_{t-1},\theta_t]+\boldsymbol{b}_o) \tag{6}$$

$$g_t=\sigma(\boldsymbol{W}_{gx}[h_{t-1},\theta_t]+\boldsymbol{b}_g) \tag{7}$$

$$c_t=f_t*c_{t-1}+i_t*g_t \tag{8}$$

$$h_t=o_t*\tanh(c_t) \tag{9}$$

其中,$\sigma(*)$是一个激活函数,$\boldsymbol{W}_{*x}$是权重矩阵,$\boldsymbol{b}_*$是偏

置向量,* 表示元素乘,h_t 是隐藏状态,包含该人在第 t 时刻的动态特征,可以模拟该人在 t 时刻所执行的动作.

将通过 LSTM 获得的个人和场景动态特征用 Z^n 和 $\hat{Z}$ 表示:

$$\hat{Z} = \mathrm{LSTM}(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_T) \tag{10}$$

$$Z^n = \mathrm{LSTM}(\theta_1^n, \theta_2^n, \cdots, \theta_T^n) \tag{11}$$

3.3　关键人物及其建模

3.3.1　关键人物定义

本文的核心思想是通过关键人物进行组群行为分析,避免无关人员的影响,通过分析,本文认为在活动场景中随时间变化具有稳定运动的人员为"关键人物". 图 2 展示了排球数据集中的"Left-spike"组群行为,标"红星"的个体为关键人物,从图中可以看出其从排球场的左侧($t-n$ 时刻)一直稳定移动到了排球场的中间($t+n$ 时刻)并且做出了击球动作,利用其个人运动和与周围相关人组成一个子集群体完成了组群行为. 图中用绿色框标出的为"非关键人物",其在整个视频中没有随时间进行稳定的运动,即,其在组群行为识别中并未做出突出贡献. 因此,本文认为确定场景中的关键人物,并且根据和关键人物相关的子群体进行组群行为识别是一种更高效的方法.

图2　排球比赛中的"Left-spike"

3.3.2　关键人物建模

如图 3 所示,根据曲线图可以看出球员从初始时刻到结束一直进行着稳定的运动. 本文堆叠所有个体的光流图像,计算其平均运动强度,进行降值排序,对组群内的所有个体赋予初始权重并且输入到 GFU 单元进行迭代训练.

通过对排球数据集分析与实验,每个组群行为由 12 个人共同完成,将运动强度排在前 5 位的个体作为关键人物可以达到最优的识别精度.

使用高精度光流[18]估计的方法进行光流特征提取,在给定的 T 帧视频中,每一帧的分辨率是 $w*h$,在第 t 帧中分别使用 $d_t^x(u,v)$ 和 $d_t^y(u,v)$ 表示在点 (u,v) $(u=1,2,3,\cdots,w;\ v=1,2,3,\cdots,h)$ 处的水平和垂直位移矢量,按如下方式堆叠连续帧的光流向量:

$$F^k(u,v,2i-1) = d_t^x(u,v) \tag{12}$$

$$F^k(u,v,2i) = d_t^y(u,v) \tag{13}$$

其中,$F^k(u,v,c)$ $(c=1,2,3,\cdots,2T)$ 表示在一个完整的 T 帧视频序列上点 (u,v) 处第 k 个人的光流特征.

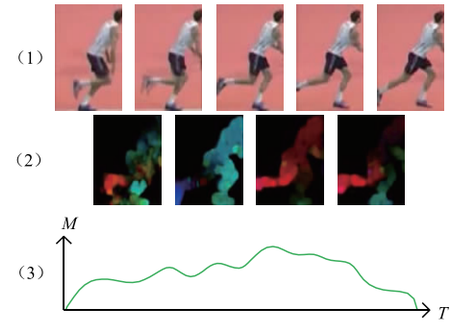

图3　关键人物建模示意图

注:第(1)行表示组群活动中随时间运动的个人图像;第(2)行是提取的光流特征热度图,其中颜色由蓝色向红色过渡,颜色越深代表运动强度越大;第(3)行当中曲线图的横轴T代表了时间,纵轴M(Motion)代表运动强度

第 k 个人的运动强度定义如下:

$$M_t^k = \frac{\sum_{u=0}^{w}\sum_{v=0}^{h}\sum_{c=2t-1}^{2t} |F^k(u,v,c)|}{w*h} \tag{14}$$

$$M^k = \left(\sum_{t=1}^{T} M_t^k\right)/T \tag{15}$$

其中,M_t^k 表示第 k 个人在第 t 帧的运动强度,M^k 表示第 k 个人在整个视频中的平均运动强度. M^k 越大,表明该个体有着长时间的稳定运动,对组群行为识别所作出的贡献也就越大. 按照 M^k 的值对个人动态特征进行降值排序,作为门控融合单元的输入.

3.4　以关键人物为核心的组群信息融合

行为识别需要用到特征融合[19]技术,文献[20]提出门控多模态单元,本文受此启发,使用最大池化策略代替级联融合,同时,接收场景特征确定个体空间位置信息,基于门控思想来选择哪个人员更有可能为正确识别组群行为做出贡献,对不同的人员赋予不同的权重. 对于组群问题来说这可以有效的降低特征维度,增强交互关系学习能力,缩短模型训练的时间,在本文中将其称为门控融合单元(Gate Fusion Unit,GFU),如图 4 所示,给出了其具体结构图.

核心组群特征融合的具体实施细节如下:

(1)门控融合单元接收关键人物建模后的动态特征 Z^n 作为输入:

$$h^n = \tanh(\boldsymbol{W}^n Z^n) \tag{16}$$

其中,$\boldsymbol{W}^n$ 是编码的权重向量,h^n 是经过编码后的个人动态特征. 关键人物建模可以帮助训练门控神经元,平均运动强度的大小决定了其初始权重的大小.

(2)以关键人物为核心的子集信息集成.

使用 sigmoid 函数设计门神经元,用符号 σ 表示,其作用是控制个体对组群行为识别所作出贡献的大小,考虑到全局特性,与第 n 个人连接的门神经元会接收所有人的动态特征以及场景动态特征作为输入,从而确

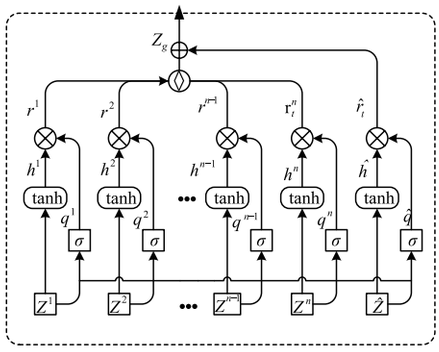

图4 门控融合单元结构图，Z^n为关键人物建模后的个人特征序列，$\hat{Z}$为场景特征，σ为门控神经元

定第 n 个人的门控输出 q^n：

$$q^n=\sigma(\boldsymbol{W}^n[Z^1,Z^2,\cdots,Z^N,\hat{Z}]) \tag{17}$$

当个体的行为与组群行为更加一致时，与其位置相近的几个成员组成了一个相关的子集群体共同完成组群活动. 本文将其之间的位置信息作为交互信息，根据子集群体的行为确定组群行为类别. 子集群体是以 q^n 为核心坐标，由关键人物特征和相关个体特征共同构成(本文规定子集群体包括 3 个个体).

(3)多子集特征融合获得组群整体信息描述.

将个人动态特征 h^n 和门控输出 q^n 进行相乘，利用场景特征中的背景信息确定个体空间位置信息，表示人与人之间的交互关系：

$$r^n=h^n\times q^n \tag{18}$$

其中 r^n 表示第 n 个人的特征输出. 将关键人物的特征和门控输出所代表的交互信息进行聚合可以有效避免无关人员对组群行为识别的影响.

同理，将场景动态特征 $\hat{Z}$ 作为输入，生成场景级的门控输出 $\hat{r}$. r^n 当中包括了个人的特征和与组群当中其他人员的交互特征，然后，本文使用最大池化策略融合所有人的特征生成组群级特征 Z_g 作为第二层 LSTM 网络的输入：

$$Z_g=r^1\diamond r^2\diamond\cdots\diamond r^n\oplus\hat{r} \tag{19}$$

其中，$\diamond$表示最大池化，场景特征不参与池化，以免影响个人特征，而是将其与核心成员子集最大池化的结果进行级联，用$\oplus$表示.

3.5 组群行为识别

本文使用第二层 LSTM 对组群级别的特征进行建模. 具体来说，通过门控融合单元后可以获得精简的整体组群行为的特征描述，而第二层 LSTM 对其进行迭代训练，旨在进一步对整体组群行为随时间的演化动态进行时序信息的刻画. 其结构设计类似于前文提到的第一层 LSTM 网络，最后连接 softmax 层进行组群行为分类，如下：

$$y=\text{softmax}(Z_g) \tag{20}$$

最终，y 即是组群行为类别. 为了说明第二层 LSTM 能够有效的模拟组群级别的时间动态，本文在 4.3.1 节中设计了基线实验说明其有效性.

4 算法验证

本文模型在公开的排球数据集上进行实验并且与现有的模型进行对比. 接下来在 4.1 节中详细介绍本文所使用的数据集并且介绍数据集的拆分训练；在 4.2 节中详细介绍实验参数设置；4.3 节对基线实验和其他模型进行了对比.

4.1 组群行为数据集

实验中采用的组群行为识别数据集是排球数据集，由 55 个视频组成，包含 4830 个注释帧. 此数据集有 9 类个人行为(action)标签：waiting，setting，digging，falling，spiking，blocking，jumping，moving，standing. 8 种组群行为标签，即每帧活动中 N 个人共同完成的场景标签：right set，right spike，right pass，right winpoint，left winpoint，left pass，left spike，left set. 如图 5，展示了排球比赛中“Left Spike”的例子，每个人都有一个行为标签，每帧图像都有一个场景活动标签.

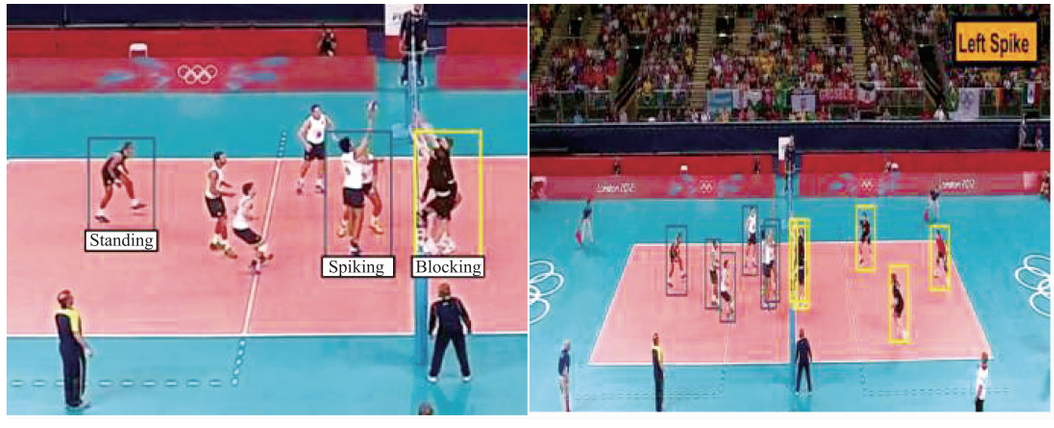

图5 排球数据集示例图

为了与现有文献公平地实验对比,使用了和文献[1]相同的训练/测试分裂方式,其中 2/3 用于训练,1/3用于测试与验证. 并且使用多类分类准确度(MCA)和平均分类准确度(MPCA)作为性能指标.

4.2　实验参数配置

本文实验基于排球数据集,实验中硬件以及软件环境如下:深度学习服务器使用 Intel Core i7-5960X(主频为 3.0GHz),图形加速使用 Nvidia GPU(型号为 GeForce GTX 1080 8G),Linux(Ubuntu 18.04)操作系统和 Pytorch1.0 深度学习框架,编程语言使用 Python 2.7.

网络模型以 RGB 图像作为输入,将图像大小调整为 224 × 224,使用目标跟踪器跟踪检测,使用在 ImageNet 预训练的 ResNet-50 网络提取静态特征. 第一层由 $N+1$ 个 LSTM 层构成,即,每人 1 个 LSTM 加上 1 个场景级的 LSTM,本文设定 $N=12$. 本文训练方式和文献[1]类似,使用端到端的方法训练 CNN + LSTM 组成的网络,LSTM 的输出通过门控融合单元映射人级和场景级之间的对应关系. 对于排球数据集,本文将 FC(K)的维度设置为 8,因为组群行为类别有 8 类.

使用 Adam[21]算法最小化成本函数,所有网络的学习率为 0.001,并且学习率在每 10 次迭代之后降低到原始值得 1/10.

4.3　实验结果分析与其他算法的对比

4.3.1　基线(Baseline)实验设计

为了验证本文模型的有效性,本文设计了 3 组 Baseline 方法进行对比.

Baseline1(B1):仅使用预先训练好的 Resnet-50 网络进行个人特征提取,使用这些个人特征级联训练 softmax 分类器. 该基线仅使用了个人的静态特征,不考虑时间动态,以说明时间动态的重要性.

Baseline2(B2):在 B1 的基础上增加了动态特征建模,使用 GFU 单元对个体特征进行融合表示组群特征连接第二层 LSTM 网络训练 softmax 分类器. 该基线旨在证明动态特征以及第二层 LSTM 网络对组群行为识别起到重要作用.

Baseline3(B3):该基线是对本文所提方法的一个简单变形,省略了关键人物建模的步骤,该基线旨在说明本文以关键人物为核心的组群行为识别方法是高效的.

4.3.2　Baseline 方法和本文算法的实验结果对比

表 1 展示了所提出模型与基线实验的对比结果,与所有的基线方法相比,本文提出的模型同时实现了最佳的 MCA 和 MPCA.

B1 方法使用了深度学习的方法提取单人静态特征并简单级联起来作为组群特征进行组群行为识别,MCA 和 MPCA 分别达到了 76.2% 和 74.6%. B2 和 B1 相比,MCA 和 MPCA 分别提升了 3.5% 和 3.1%,说明了时间动态对组群行为识别的重要性. 通过 B3 和 B2 相比,实验结果得到了进一步的提升,说明了所使用的分层模型和 GFU 起到了关键性的作用. 通过本文的方法与 B3 相比,可以发现关键人物建模的作用,将 MCA 和 MPCA 分别提高了 1.9% 和 2.3%,说明了通过关键人物进行组群行为识别的可行性和有效性. 并且在实验中,本文发现,将关键人物建模输入到 GFU 后,GFU 的训练时间减少了 22min. 通过以上对比,本文的模型获得了更好的性能.

表 1　在排球数据集上与基线方法的比较

Approach	MCA	MPCA
Baseline1	76.2%	74.6%
Baseline2	79.7%	77.7%
Baseline3	83.5%	84.4%
Ours	85.4%	86.7%

4.3.3　本文算法和现有模型的实验结果对比

表 2 展示了所提模型与现有模型的结果对比,第一组结果(1 group)是将排球比赛中两队人员视为一组;第二组(2 group)是将两队人员分成两组并分别从中提取特征. 根据表中的结果观察到,相对于将图片人员视为 1 组,将球员分为 2 组时,识别精度明显提高. "1 group"的方法是同时提取所有人的特征,"2 group"的方法是首先使用真实标签注释将球员分为相对应的 2 个子团体,然后分别在这 2 个子团体中提取特征,最后将两组特征汇集表示组群特征. 但是,当真实标签不容易获得的时候,这种分割为 2 组的方法就是一个额外的开销.

表 2　在排球数据集上与现有先进结果的对比

Model	MCA	MPCA
2-layer LSTMs[1] (1group)	70.3%	65.9%
CERN[4] (1 group)	73.5%	72.2%
SRNN[5] (1 group)	73.39%	NA
2-layer LSTMs[1] (2group)	81.9%	82.9%
CERN[4] (2 group)	83.3%	83.6%
SRNN[5] (2group)	83.47%	NA
OURS	85.4%	86.7%

相比之下,本文提出的模型接收所有个体特征并使用门控融合单元(GFU)学习每个人对组群行为识别所做出的贡献,结果优于"1 group"和"2 group"的方法,通过实验证明,这是 GFU 用于特征融合过程中增强了结构学习所实现的. 本文考虑了在特定的时间步骤中的场景特征和所有人的动态特征,这就使得模型可以根据场景上下文信息有效地改变对每个人特征的关注

程度.

文献[4]和文献[5]通过 LSTM 对时间进行建模,并实现了性能上的提升,首先,训练一个单人级 LSTM 对场景中的每个人生成个体动作概率分布,在组群级别的 LSTM 上利用这些分布决定组群行为,没有使用人级特征和个人位置相关的场景结构信息. 相比之下,通过利用人级和场景级特征能够对模型性能进行提升,实验结果证明本文提出将场景特征作为输入起到了重要的作用.

5 总结

本文提出了一种以关键人物为核心的组群行为识别模型,通过预训练的 CNN 网络进行静态特征提取,然后使用第一层 LSTM 提取底层动态特征(个人和场景),并且对关键人物按照运动强度进行建模,再经过门控融合单元进行特征融合代表高级特征(组群特征)输入到第二层 LSTM 并且连接 softmax 分类器进行分类输出,达到组群行为识别的目的.

通过对比本文所提出模型与基线方法以及现有先进方法的实验结果,显示出了本文方法的优越性,主要表现在对于组群行为内部的结构学习能力,模型训练时间以及上下文信息的处理等问题上. 证明了门控融合单元、关键人物建模以及分层 LSTM 在组群行为识别任务中起到了重要作用,同时也证明了本文所提模型的有效性以及合理性. 接下来,本文计划优化特征提取过程,微调深度网络,减少模型训练的消耗,并且使模型能够应用于更多的数据集.

参考文献

[1] Ibrahim M S, Muralidharan S, Deng Z, et al. A hierarchical deep temporal model for group activity recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2016. 1971 – 1980.

[2] Jain A, Zamir A R, Savarese S, et al. Structural-rnn: Deep learning on spatio-temporal graphs[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 5308 – 5317.

[3] Deng Z, Vahdat A, Hu H, et al. Structure inference machines: recurrent neural networks for analyzing relations in group activity recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2016. 4772 – 4781.

[4] Shu T, Todorovic S, Zhu S C. CERN: Confidence-energy recurrent network for group activity recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2017. 4255 – 4263.

[5] Biswas S, Gall J. Structural recurrent neural network (srnn) for group activity analysis[A]. IEEE Winter Conference on Applications of Computer Vision (WACV)[C]. USA: IEEE, 2018. 1625 – 1632.

[6] Gammulle H, Denman S, Sridharan S, et al. Multi-Level Sequence GAN for Group Activity Recognition[A]. Asian Conference on Computer Vision[C]. Cham: Springer, 2018. 331 – 346.

[7] SHI Lei, ZHANG Yifan, CHENG Jian, LU Hanqing. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2019. 12026 – 12035.

[8] Wu J, Wang L, Wang L, et al. Learning actor relation graphs for group activity recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2019. 9964 – 9974.

[9] Ibrahim MS, Mori G. Hierarchical relational networks for group activity recognition and retrieval[A]. European Conference on Computer Vision[C]. USA: ECCV, 2018. 721 – 736.

[10] Yan R, Tang J, Shu X, et al. Participation-contributed temporal dynamic model for group activity recognition[A]. ACM Multimedia Conference[C]. USA: ACM, 2018. 1292 – 1300.

[11] Kong L, Qin J, Huang D, et al. Hierarchical attention and context modeling for group activity recognition[A]. Proceedings of ICASSP[C]. USA: IEEE, 2018. 1328 – 1332.

[12] 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.

[13] 桑海峰,王传正,吕应宇,何大阔,刘晴. 基于多信息流动卷积神经网络的行人再识别[J]. 电子学报,2019,47(2):351 – 357.
SANG Hai-feng, WANG Chuan-zheng, LÜ Ying-yu, HE Da-kuo, LIU Qing. Person re-identification based on multi-information flow convolutional neural network[J]. Acta Electronica Sinica, 2019, 47(2): 351 – 357. (in Chinese)

[14] Danelljan M, Khan F, Felsberg M, et al. Accurate scale estimation for robust visual tracking[A]. British Machine Vision Conference, Nottingham[C]. [S. L.]: BMVA, 2014. 1 – 5.

[15] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211 – 52.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition[C]. USA: IEEE, 2016. 770 – 778.

[17] 吕品,李全刚,柳厅文,宁振虎,王玉斌,时金桥,方滨兴. 基于双向 LSTM 的误植域名滥用检测方法[J]. 电子学报,2018,46(9):2081 – 2086.
LÜ Pin, LI Quan-gang, LIU Ting-wen, NING Zhen-hu,

WANG Yu-bin, SHI Jin-qiao, FANG Bin-xing. Towards typo squatting abuse detection using bi-directional LSTM [J]. Acta Electronica Sinica, 2018, 46(9): 2081 – 2086. (in Chinese)

[18] Brox T, Bruhn A, Weickert J, et al. High accuracy optical flow estimation based on a theory for warping[A]. European Conference on Computer Vision[C]. Berlin: Springer-Verlag, 2004. 25 – 36.

[19] 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)

[20] Arevalo J, Solorio T, et al. Gated Multimodal Units for Information Fusion[M]. London: ICLR, 2017. 0941 – 0643.

[21] Kingma D P, Ba J. Adam: A method for stochastic optimization[A]. International Conference on Learning Representations (ICLR)[C]. Ithaca, NY: arXiv. org, 2014. 13.

作者简介

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 年分别在石油大学(华东)、石油大学(北京)工业自动化和中国海洋大学获应用电子技术学士、硕士学位和博士学位. 主要从事计算机视觉方面的有关研究.
E-mail: Wangchuanxu_qd@163.com

薛　豪　男,1994 年 1 月出生,山东临沂人. 2016 年毕业于潍坊学院信息与控制工程学院,取得学士学位,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: xuehao1130@yeah.net