文章编号：2095-2163(2022)09-0045-06 中图分类号：TP391 文献标志码：A

融合时间和空间上下文特征的群体行为识别

李 骏[1]，程雅儒[1]，谢 昭[1]，孙永宣[1]，吴克伟[1,2]，武金金[1]
（1 合肥工业大学 计算机与信息学院，合肥 230601；2 合肥工业大学 工业安全与应急技术安徽省重点实验室，合肥 230601）

摘 要：群体行为识别任务中，行为特征具有复杂的时空特性。为了实现有效的行为特征时间编码，本文提出一种融合时间和空间上下文特征的群体行为识别模型。为了分析个体行为特征的时间上下文依赖关系，设计了通道级时间上下文模块，该模块对个体特征的多个通道进行时间平移；分别研究时间延迟移动、时间双向移动、时间循环双向移动的 3 种策略，并讨论各种策略下通道比例对时间上下文估计的作用。其次，构建了基于融合通道级时间上下文特征的空间图模型，用于对个体空间上下文的编码。该模型使用外观和位置估计初步的个体之间的空间上下文关系，并进一步设计多图策略，来估计多种可能的个体之间的关系。最后，对图模型编码的个体特征，使用个体池化获得群体特征，并使用多层感知器来识别群体行为。本文方法在 Volleyball 和 Collective Activity 数据集上优于现有群体行为识别方法，设计的时间上下文特征具有良好个体行为编码能力。

关键词：群体行为识别；时间上下文；时间位移策略；空间上下文；多图模型

Group activity recognition based on temporal and spatial context features

LI Jun[1], CHENG Yaru[1], XIE Zhao[1], SUN Yongxuan[1], WU Kewei[1,2], WU Jinjin[1]
(1 School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China; 2 Anhui Province Key Laboratory of Industry Safety and Emergency Technology, Hefei University of Technology, Hefei 230601, China)

【Abstract】In a group activity, individuals have complex spatial-temporal features. To encode the complex spatial-temporal features, the paper proposes a group activity recognition model based on temporal and spatial context features. First, to analyze the temporal context-dependency in individual features, a channel-wise temporal context module is designed, which uses a shift strategy to learn temporal context. Three strategies are studied, including temporal delay shift, temporal bi-direction shift, and temporal recurrent bi-direction shift, and the shift ratio in the shift strategy is also discussed. Second, a spatial graph model based on fusing channel-level temporal context features is constructed to encode the spatial context of the individual. The initial spatial context relation is estimated with both appearance feature and position feature. Furtherly multiple graph strategy is used to represent multiple relations. Finally, temporal pooling is used to aggregate the individual features into group features and multiple layer perceptron is used to predict the group activity. Experimental results in the Volleyball dataset and the Collective Activity dataset show that the proposed method outperforms the state-of-the-art methods. The proposed temporal context features encode the individual features well.

【Key words】group activity recognition; temporal context; temporal shift strategy; spatial context; model with multiple graphs

0 引 言

群体行为识别，是通过对人员密集场所的视频分析，并对其突发性群体行为进行识别，有利于维护公共场所安全，避免人员伤亡和财产损失，已被广泛应用于视频监控、视频摘要、视频检索等领域。个体行为识别模型只需要识别个体的单独行动，而群体行为识别，需要依据个体的行为，推断出个体之间的群体活动。视频中，个体的关系是隐藏的，且行为特征具有复杂的时序信息，个体之间的行为会相互干扰，影响多人关系的估计结果，而解析个体的时序信息具有一定的挑战性。

群体的外观特征通常使用卷积神经网络来提取，但无法提取群体的时序信息。实验表明，虽然可

基金项目：安徽省重点研究与开发计划(202004d07020004)；安徽省自然科学基金项目(2108085MF203)；中央高校基本科研业务费专项资金(PA2021GDSK0072, JZ2021HGQA0219)。

作者简介：李 骏(1996-)，男，硕士研究生，主要研究方向：计算机视觉、人工智能、模式识别；程雅儒(2001-)，女，本科生，主要研究方向：计算机视觉；谢 昭(1980-)，男，博士，副研究员，主要研究方向：计算机视觉、图像分析与理解、模式识别；孙永宣(1980-)，男，博士，副研究员，主要研究方向：计算机视觉、图像分析与理解、模式识别；吴克伟(1984-)，男，博士，副研究员，主要研究方向：计算机视觉、人工智能、模式识别；武金金(1996-)，女，本科生，主要研究方向：计算机视觉。

通讯作者：吴克伟 Email：wu_kewei1984@163.com

收稿日期：2022-03-06

以利用长短期记忆网络(LSTM)提取个体的时序信息,但会导致网络性能下降。现有的图模型结构只专注于群体的外观信息和位置信息,不能够很好地表达群体关系,导致群体行为识别效果欠佳。

针对上述问题,本文提出了一种时间上下文模块,用来解决个体特征缺乏时序信息的问题。通过通道级的时间位移方法,每个个体的时序信息都得到增强。为了保证群体建模的完整性,构建了基于融合通道级时间上下文特征的空间图模型,该图模型使用外观和位置信息,实现对空间关系的编码。在增强时序信息的基础上,通过建立多个个体关系图来模拟个体之间的相互关系,将每个个体的全部特征描述为图模型的每个节点,通过图模型的推理,完成行为分类。

1 相关工作

1.1 视频特征学习

早期的视频特征学习主要采用传统手工制作的视觉特征[1],或采用与概率图模型结合的方法[2]。在图模型的基础上,多尺度模型 And-or[3] 通过对不同的群体粒度进行建模,对群组行为分类。双流卷积神经网络[2] 还可以额外学习视频帧的光流图像特征,进一步识别不同的行为。时间分段网络[4] 在双流的基础上做出改进,通过稀疏采样和加权池化来识别行为特征。膨胀三维卷积网络[5] 通过将 2D CNN 参数膨胀拓展为 3D CNN,可以解决 TSN 单一视频权重的问题。

1.2 交互关系分析

群体行为分析的细节存在于群体结构中。与个体行为识别不同,群体行为识别更重要的是分析个体之间交互关系。层次关系网络(HRN)[6] 使用固定的群体结构,来学习个体之间的相互关系强度。卷积关系机(CRM)[7] 使用多阶段的群体结构误差,来优化群体行为识别结果。时空注意力图网络 stagNet[8] 被用于估计图结构中,用于表达目标之间的关系。

图卷积网络(GCN)[9] 在结构化数据的表示和推理方面具有优势。图注意力交互模型(GAIM)[10] 将群体节点加入图模型,并利用自注意力同时学习个体之间和个体与群体之间的关系。在图模型中引入 LSTM 可以增强时序信息。置信度能量循环网络(CERN)[11] 在 LSTM 的动态特征基础上构建图模型,在图模型构建阶段,可以获得群体的时序信息。本文在模型的设计中应用了图卷积网络[12],将个体的信息作为图模型一个节点。为了保证群体建模的完整性,在图构建的过程中引入了多图策略。

2 融合时间和空间上下文特征的群体行为识别

本文使用 Inception-v3[12] 对视频序列提取特征,通过 RoIAlign[13] 从帧特征图中提取每个个体的边界框特征,将对齐的特征通过全连接层得到每个个体的原始特征。原始特征经过通道级时间上下文模块,与图卷积特征相加得到多图融合特征,最终融合特征通过群体分类器和个体分类器完成行为的分类。整体网络框架如图 1 所示。

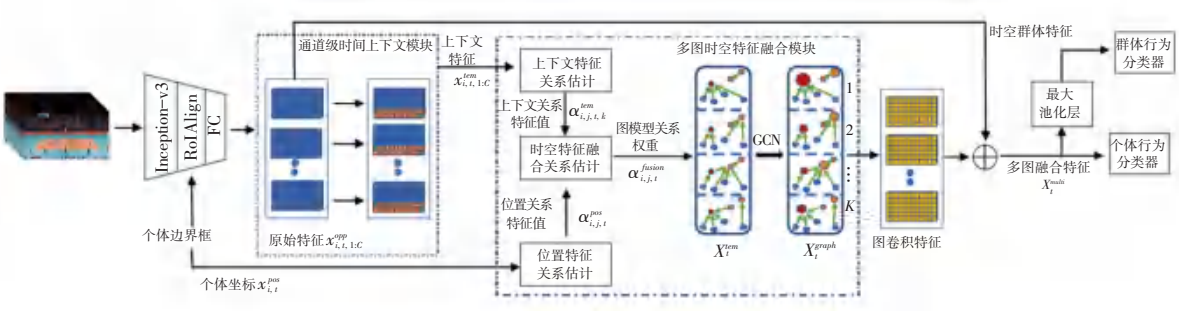


图 1 融合时间和空间上下文特征的群体行为识别网络

Fig. 1 The group activity recognition model based on temporal and spatial context features

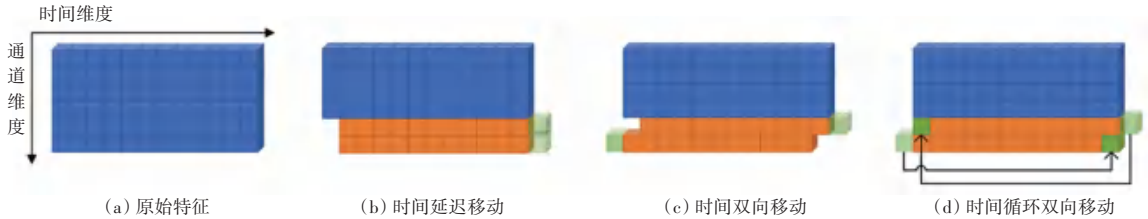
2.1 通道级时间上下文模块

本文设计了通道级时间上下文模块,该模块通过对个体特征的多个通道进行时间平移,可以让视频帧获得相邻帧的时序信息,在图模型的建立过程中增强模型的时序信息,并最终影响行为分类的结果。

通道级位移策略如图 2 所示，对于个体特征的通道位移，本文分别采用时间延迟后移、时间双向移动、时间循环双向移动策略来实现。

图 2 中描述了本文设计的 3 种位移方式，考虑了不同的位移方式对于模型性能的影响，并最终选择时间循环双向移动作为模块内特征位移的方式。

通过时间循环双向移动的位移策略，既增强了时序信息，也确保个体特征不会丢失，保证了图模型构建过程中建模的完整性。



(a) 原始特征　(b) 时间延迟移动　(c) 时间双向移动　(d) 时间循环双向移动

图 2　通道级位移策略

Fig. 2　Channel-wise shift strategies

2.2　多图时空特征融合模块

由于图模型能够实现结构化数据的表示和推理，本文在建模中利用图模型来模拟群体行为中的成对个体关系。图定义为：$G=\{V,E\}$，其中，节点 $V=\{v_i\}$，边 $E=\{e_{i,j}\}$，节点编号为 $i=1,2,\cdots,N$，$j=1,2,\cdots,N$，这里 N 表示群体中个体的数量；节点 V 有外观特征和位置特征；E 表示图模型节点之间的相互关系。通过估计边上的关系取值，构成关系矩阵 $\boldsymbol{A}$，表示个体 i 和个体 j 的关联性。

在考虑上下文建模时，对 2 个个体的特征使用线性变换来学习投影特征，在投影的基础上，通过点积和归一化来估计 2 个个体的关系。使用 $\alpha_{i,j,t,k}^{tem}$ 来表示学习到的 2 个个体上下文特征关系值，计算方式如下：

$$\alpha_{i,j,t,k}^{tem}=\frac{(w_k^1 x_{i,t}^{tem})^{\mathrm{T}} w_k^2 x_{j,t}^{tem}}{\sqrt{d_x}} \tag{1}$$

其中，w_k^1 和 w_k^2 分别表示不同外观特征的线性变换参数（$k=1,2\ldots,K$ 为多图的编号），可以通过不同的外观相似度，来生成多个图；$x_{i,t}^{tem}$ 为个体的上下文特征；d_x 是 $x_{i,t}^{tem}$ 的维度，作为归一化参数。

在群体活动中，位置较近的 2 个个体更容易发生交互行为。在推导个体之间位置特征关系时，设定一个阈值，当 2 个个体的距离小于阈值时，将会计算个体之间的位置特征关系。若使用 $\alpha_{i,j,t}^{pos}$ 表示 2 个个体的位置特征关系值，对此则可表示为：

$$\alpha_{i,j,t}^{pos}=I(dist(x_{i,t}^{pos},x_{j,t}^{pos})\leqslant u) \tag{2}$$

其中，$x_{i,t}^{pos}$、$x_{j,t}^{pos}$ 分别表示个体 i 和个体 j 在视频帧中的坐标；$dist(x_{i,t}^{pos},x_{j,t}^{pos})$ 表示 2 个 i 和 j 之间的欧式距离；$I(\cdot)$ 是指示函数；u 为阈值参数，如果 2 个个体距离小于 u，将 $\alpha_{i,j,t}^{pos}$ 置为 1，否则为 0。

对于给定的上下文特征关系值和位置特征关系值，使用线性变化融合成一个权重，并将每个个体的权重归一化。使用 $\alpha_{i,j,t,k}^{fusion}$ 表示归一化后的个体关系权重，推得的数学定义式可写为：

$$\alpha_{i,j,t,k}^{fusion}=\frac{\alpha_{i,j,t}^{pos}\cdot\exp(\alpha_{i,j,t,k}^{tem})}{\sum_{j=1}^{N}\alpha_{i,j,t}^{pos}\cdot\exp(\alpha_{i,j,t,k}^{tem})} \tag{3}$$

本文建立了一组多图的关系矩阵进行图推理。使用图卷积网络实现了图的推理过程，对于图中的目标节点，根据其周围全部个体的权重进行更新。研究中使用 X_t^{graph} 来表示图模型输出的特征，其数学表述见如下：

$$X_t^{graph}=\sum_{k=1}^{K}\sigma(\boldsymbol{A}_{t,k}\boldsymbol{X}_t^{tem}\boldsymbol{W}_k) \tag{4}$$

其中，$\boldsymbol{A}_{t,k}$ 表示个体关系权重；$\boldsymbol{X}_t^{tem}$ 表示时间上下文特征；$\boldsymbol{W}_k$ 表示权重参数矩阵；$\sigma(\cdot)$ 是非线性激活函数。将 K 张图推理得到的特征，以对应维度相加的方式进行融合，作为图模型的输出特征。

将图模型输出的特征与原始特征融合相加，得到多图融合特征 X_t^{multi}，即：

$$X_t^{multi}=X_t^{graph}+X_t^{app} \tag{5}$$

2.3　群体行为识别

将多图融合特征 X_t^{multi} 与权重参数矩阵做线性变换，可以得到每一帧的结果，并将视频序列的平均预测结果作为个体行为识别的结果。使用 y_i^p 表示行为预测标签，则可推得：

$$y_i^p=\frac{1}{T}\sum_{t=1}^{T}w^p x_{i,t}^{multi} \tag{6}$$

将多图融合特征通过 Max Pooling 池化层减少维度，得到群体行为特征。并将群体行为特征与权重参数矩阵做线性变化，可以得到每一帧的结果，将视频序列的平均预测结果作为群体行为识别的结果。群体行为的预测标签 y^G 数学计算公式具体如

下：

$$y^G = \frac{1}{T}\sum_{t=1}^{T} w^G maxpooling_i(x_{i,t}^{multi}) \quad (7)$$

2.4 损失函数

整个模型可以通过反向传播的方式，进行端到端的训练，使用损失函数来评价预测值和真实值偏差的程度，损失函数的运算公式可写为：

$$Loss = L(y^G, y_{gt}^G) + L(y^P, y_{gt}^P) \quad (8)$$

其中，$L(\cdot)$ 是交叉熵损失，y_{gt}^G 和 y_{gt}^P 是群体活动和个体活动的真实标签，使用分类器得到群体活动的预测 y^G 和个体活动的预测 y^P。式中第一项代表群体活动分类损失，第二项代表个体活动的分类损失。

3 实验

3.1 数据集与评价标准

本文在 Volleyball 数据集[6]和 Collective Activity 数据集[13]上分别进行了实验。对此拟做阐释分述如下。

(1) Volleyball 数据集。由 55 场排球比赛中收集的 4 830 个视频片段组成，其中包括 3 493 个训练片段，1 377个测试片段。在每个视频片段中，视频的中间帧标注了个体的边界框、个体行为标签和群体行为标签。总地说来，群体行为标签有 8 种，分别是 Right set、Right spike、Right pass、Right winpoint、Left set、Left spike、Left pass、Left winpoint；个体行为标签有 9 种，分别是 Blocking、Digging、Falling、Jumping、Moving、Setting、Spiking、Standing、Waiting。实验中，使用一个长度为 $T = 10$ 的时间窗口，对应于标注帧的前 5 帧和后 4 帧。未被标注的个体边界框数据从该数据集提供的轨迹信息数据中获取。

(2) Collective Activity 数据集。由低分辨率相机拍摄的 44 个视频片段组成，总共约为2 500帧。每个视频片段每 10 帧有一个标注，标注包含个体行为和群体行为标签，以及个体的边界框。共 5 个群体活动标签，分别为 Crossing、Waiting、Queueing、Walking、Talking；6 个个体行为标签，分别为 NA、Crossing、Waiting、Queueing、Walking、Talking。实验中的 2/3 视频用于训练，其余用于测试。

本文采用多类正确率（Multi-Class Accuracy，*MCA*）作为评价标准，先求出所有类别的正确样本数，并除以所有类别的样本总数来获得多类正确率。

3.2 实验环境及参数设定

本文实验使用 Inception-v3[12]提取视频特征，RoIAlign[13]为每个个体提取 1 024 维度特征，这些特征是在每个个体边界框约束下提取的。数据集参数设定如下：

(1) Volleyball 数据集。网络超参设置为：*batch size* 为 8，*Dropout* 参数为 0.3，学习率初始设置为 1e-4，权重参数 u 为图片宽度的 1/5，网络训练 180 个周期，每 30 个周期学习后变为之前的 0.5 倍，学习率在 4 次衰减后停止衰减。

(2) 对于 Collective Activity 数据集。网络超参设置为：*batch size* 为 16，*Dropout* 参数为 0.5，初始学习率为 1e-3，权重参数 u 为图片宽度的 1/5，网络训练 80 个周期，每 10 个周期学习率变为之前的 0.1 倍，学习率在 4 次衰减后停止衰减。

实验在 64 位 Ubuntu16.04 上进行，编程环境选择 Python3.7，实验采用 Pytorch1.4 深度学习平台。计算机配置英特尔 Xeon（R）W-2133 处理器，内存为 64 G，配有 2 块 GeForce RTX 2080Ti 显卡。

3.3 对比实验

在 Volleyball 数据集上，本文方法与其它方法对比的结果见表 1。由表 1 可以看出，本文方法的效果优于其它方法，其识别准确率相比于 VC 模型[14]提高了 1.0%。在个体行为准确率识别中，也表现出了最佳的性能，相比于 AT 模型[15]提高了 0.4%。

在 Collective Activity 数据集上，本文方法与其它方法对比的结果见表 2。由表 2 可知，本文方法性能优于现有的行为识别方法。在群体行为识别准确率上，本文模型相对于 VC 模型提高了 0.4%；在个体行为识别准确率上，相对于 GLIL 模型[16]提高了 0.2%。

表 1 在 Volleyball 数据集上与其它方法的对比

Tab. 1 Comparison with the state-of-the-art methods on Volleyball dataset

方法 Backbone	Backbone	群体行为 *MCA*/%	个体行为 *MCA*/%
HDTM（2016）	AlexNet	81.9	-
SSU（2017）	Inception-v3	90.6	81.8
CERN（2017）	VGG16	83.3	-
HRN（2018）	VGG19	89.5	-
ARG（2019）	Inception-v3	92.5	82.8
CRM（2019）	I3D	93.0	-
GAIM（2020）	Inception-v3	91.9	-
AT（2020）	I3D	93.0	83.7
stagNet（2020）	VGG16	89.3	82.3
VC（2021）	Inception-v3	93.3	-
本文方法	Inception-v3	94.3	84.1

表 2　在 Collective Activity 数据集上与其它方法的对比

Tab. 2　Comparison with the state－of－the－art methods on Collective Activity dataset

方法 Backbone	Backbone	群体行为 *MCA*/ %	个体行为 *MCA*/ %
HDTM（2016）	AlexNet	81.5	－
CERN（2017）	VGG16	87.2	88.3
ARG（2019）	Inception－v3	91.0	－
CRM（2019）	I3D	85.8	－
GAIM（2020）	Inception－v3	90.6	－
AT（2020）	I3D	92.8	－
stagNet（2020）	VGG16	89.1	－
GLIL（2021）	Inception－v3	－	94.9
VC（2021）	Inception－v3	95.1	－
本文方法	Inception－v3	95.5	95.1

3.4　消融实验

为了验证本文方法的有效性以及各个模块的效果，在 Volleyball 数据集上进行消融实验分析。设计了一种特征通道位移的时间上下文模块，讨论了通道位移策略对于识别准确率的影响。实验效果数据见表 3。

表 3　在 Volleyball 数据集上不同位移方式的效果

Tab. 3　Effects of different shift modes on Volleyball dataset

位移策略	群体行为 *MCA*/ %	个体行为 *MCA* /%
原始特征	91.2	80.4
时间延迟移动	90.1	79.6
时间双向移动	91.6	81.0
时间循环双向移动	92.5	81.9

由表 3 可见，在使用时间循环双向移动时，既得到完整的时序信息，也保证了个体特征的完整性，且正确率得到了明显的提升。因此，本文最终选择时间循环双向移动策略。

3.5　可视化分析

实验中使用 t－SNE 来可视化不同模型的标签分离度。其可视化结果如图 3 所示。

从图 3 中可以看出，相对于 VC 模型，本文方法在 Right pass 和 Right winpoint 这 2 类群体行为中有着更好的分离度，其它行为的分离度也优于 VC[14] 和 MLIR 模型[17]，验证了使用本文方法学习到的场景特征有更好的分离效果。

（a）MLIR　　（b）VC　　（c）本文方法

图 3　在 Volleyball 数据集上 t－SNE 可视化

Fig. 3　t－SNE visualization on Volleyball dataset

4　结束语

本文提出了一种新的通道时间上下文模块，通过在特征通道层面进行通道时间位移，使用时间循环双向移动作为位移策略，有效增强了个体的时序信息。其次，本文构建了基于融合通道级时间上下文特征的空间图模型，实现多复杂空间关系的编码。通过在 2 个公开的数据集上进行试验分析，结果显示本文方法优于现有群体行为识别方法，验证了本文方法的有效性。

参考文献

[1] WANG Limin, LI Wei, LI Wen, et al. Appearance－and－relation networks for video classification [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA:IEEE, 2017: 1430－1439.

[2] SIMONYAN K , ZISSERMAN A . Two－stream convolutional networks for action recognition in videos [C]// International Conference on Neural Information Processing Systems. Kuching, Malaysia:MIT Press, 2014:568－576.

[3] AMER M R, DAN Xie, ZHAO Mingtian , et al. Cost－sensitive top－down/bottom－up inference for multiscale activity recognition [M]// FITZGIBBON A, LAZEBNIK S, PERONA P, et al. Computer Vision－ECCV 2012. ECCV 2012. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg,2012,7575:187－200.

[4] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks for action recognition in videos [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 41(11):2740－2755.

[5] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA:IEEE, 2017:4724－4733.

[6] IBRAHIM M, MURALIDHARAN S, DENG Zhiwei, et al. A hierarchical deep temporal model for group activity recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA:IEEE,2016: 1971－1980.

（下转第 55 页）