基于多流架构与长短时记忆网络的组群行为识别方法研究

王传旭,胡小悦,孟唯佳,闫春娟
(青岛科技大学信息科学技术学院,山东青岛 266001)

摘 要: 提出一种基于多流架构与长短时记忆网络的上下文建模框架,旨在克服组群行为识别的两个难点,其一对复杂场景中多视觉线索进行信息融合;其二对情景人物进行建模,以获得长视频上下文关系.并且,对基于全局信息和基于局部信息的识别结果进行决策融合,判定最终组群行为属性.该算法在CAD1和CAD2上分别取得93.2%和95.7%平均识别率.

关键词: 组群行为识别;多视觉线索融合;交互上下文建模;全局-局部模型;长短时记忆网络
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112 (2020)04-0800-08
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.04.022

Research on Group Behavior Recognition Method Based on Multi-Stream Architecture and Long Short-Term Memory Network

WANG Chuan-xu, HU Xiao-yue, MENG Wei-jia, YAN Chun-juan
(*Institute of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, Shandong* 266001, *China*)

Abstract: This paper proposes a context modeling framework based on multi-stream architecture and LSTM, which aims to overcome two difficulties for group behavior recognition. One is to fuse information from multiple visual cues in complex scenes, the other is to model situational characters to get the long-term temporal context in the video. In addition, decision fusion is performed on the behavior recognition results based on global information and local information to determine the final group behavior attributes. The algorithm achieved 93.2% and 95.7% average recognition rates on CAD1 and CAD2 respectively.

Key words: group behavior recognition; fusion of multiple visual cues; interactive context modeling; global-local model; long short-term memory network

1 引言

近年来,视频中的人类行为识别[1~5]在计算机视觉领域取得了举世瞩目的成就.组群行为分析在现实生活中也得到了广泛应用,如智能视频监控、异常事件检测、社会行为理解等.目前,基于传统人工设计视觉特征[6~8]实现行为识别的方法,由于识别精度低,已经完全被深度学习方法[9~12]所取代.文献[10]提出一种可信度能量循环网络(CERN)模型,并结合两层长短时记忆网络(Long Short-Term Memory Network,LSTM)对组群行为进行预测,该模型将能量层代替普通的softmax层进行预测,当能量达到最小时,得到组群行为类别标签.文献[11]提出一种基于LSTM网络的分层模型,该模型由两层LSTM组成:第一层提取个人层次上的特征表示;第二层聚合个人特征,捕获组群级的特征表示,最终实现对组群的行为识别.文献[12]提出一种参与式时间动态模型(Participation-Contributed Temporal Dynamic Model,PC-TDM),该模型由三层LSTM组成:第一层单人LSTM旨在为每个人的个人动态建模,第二层交互Bi-LSTM旨在捕获组内人与人之间的交互信息,第三层聚合LSTM聚合单人特征和交互信息,获得组群行为标签.以上这些方法都是基于深度学习的,在组群行为方面都取得了很好的结果,但都没有将光流信息及背景信息考虑在内.

相比较单人行为识别,组群行为识别具有更复杂的结构,除了个人行为之外,还需考虑人与人,组群与人之间的交互关系,这些交互关系解决了集体行为识别中存在的模糊性问题.如图1所示,如果只考虑红色框内的“个人”,无视环境,其行为是“站立”;但考虑了周

收稿日期:2019-07-15;修回日期:2019-12-17;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61672305)

围环境之后,这个人可能正在与他人“交谈”或“排队等候”. 因此,为了正确理解组群行为,必须将所有人和场景信息一起考虑. 但在一段视频中,可能存在无关人物,对识别造成干扰,这就首先要求识别视频中的主要人物,并结合场景信息来判断组群行为类别.

此外,研究视频中复杂的组群行为,不能仅依靠外观信息,还需考虑更重要的运动信息. 因此,本文捕获了视频中人物及场景的光流信息,采用双流卷积神经网络 TSN 网络(Temporal Segment Networks)来处理该问题,并结合 LSTM 网络架构,捕获视频中长期依赖关系,以便更快速地捕获行为的上下文信息.

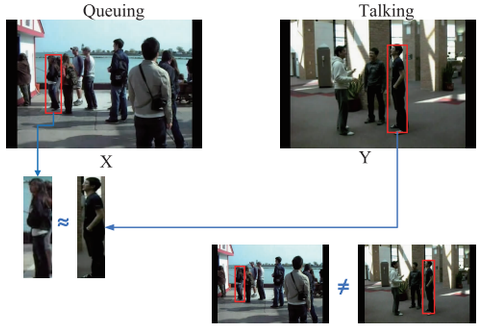

图1　组群行为识别存在的问题,个人行为大致相似,但组群行为不同

2　相关工作

近年来,双流网络的应用也受到重视. 文献[13]最早提出双流网络,通过在光流上训练另一个神经网络来整合运动信息,利用空间特征和光流特征进行初步行为识别,结合融合概率分数,使行为识别准确性得到显著提升. 但其空间网输入是连续的 RGB 视频帧,时间网输入是连续的光流信息,该双流网络存在两方面问题:一是其输入为连续视频帧,存在高度冗余;二是组群行为相对于单人行为更具有复杂性和长时间依存关系,因而其不适合对组群行为进行长时间建模. 此后,许多研究学者在此双流网络的基础上进行了改进[14,15],在单人行为识别方面都取得了较好的结果. 因此本文最终采用 TSN 网络提取视频中的外观及运动信息,并将该框架扩展到处理集体行为识别问题上.

目前,越来越多的方法[16~19]使用深度学习网络对组群行为识别进行研究,并取得了不错成效. 因此本文基于以上类似想法,提出一种基于 TSN 网络与 LSTM 网络相结合的架构,不仅考虑主要人物的时空特征,还将背景信息考虑在内,从而获得全局和局部的外观及光流信息,通过级联融合策略以合并来自多个通道的信息,连接 LSTM 网络,得到组群级特征表示,最终实现组群行为识别,并取得了不错的效果.

3　算法描述

多重视觉线索在行为识别中起着愈发重要的作用,受 TSN 网络启发,本文主要贡献是利用视频中的多重视觉线索,不仅考虑外观特征,还将运动特征考虑在内;在关注局部信息的同时,更加关注全局特征的有效性. 算法模型如图 2 所示,概述如下.

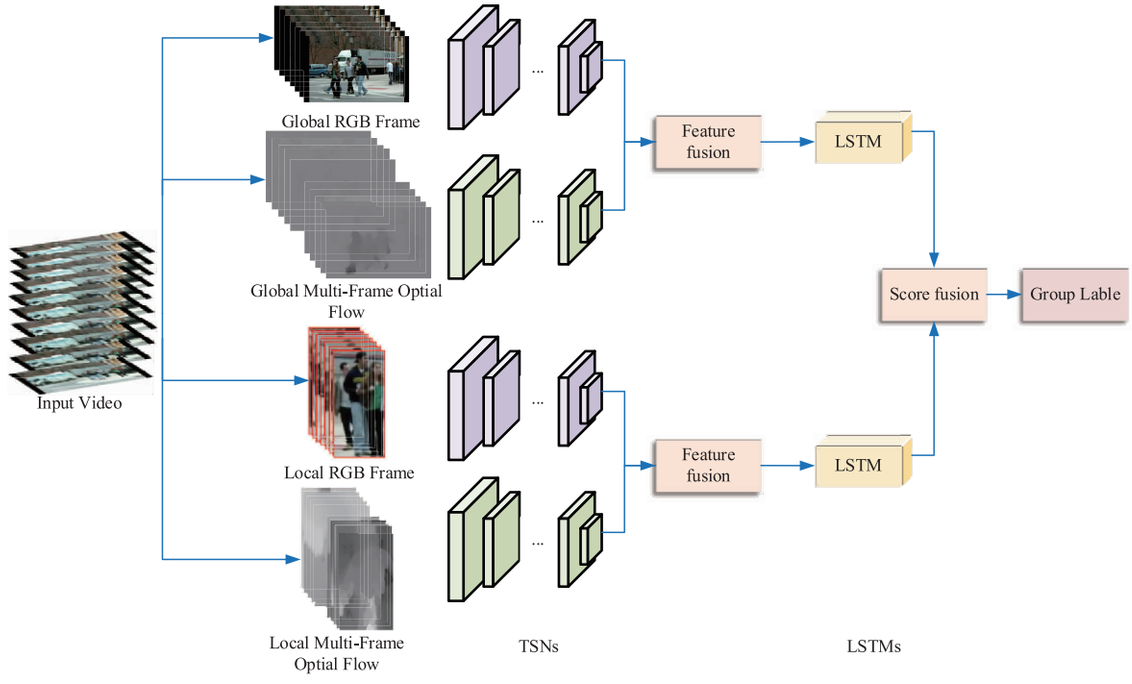

图2　本文整体算法架构

首先,利用数据集中提供的人体目标候选框进行特征提取,得到局部空间特征表示. 其次,利用 TV-L1(Total Variation-L1)光流算法分别提取视频中整幅图片和主要人物的光流时序信息,即短期运动特征. 然后,一方面将整幅 RGB 图像和其对应的光流图像分别输入到全局 TSN_G 空间网和时间网中提取全局时空特征表示;另一方面将带有候选框的局部 RGB 图像和其对应的局部光流图像分别输入到局部 TSN_L 空间网和时间网中,提取人体目标的局部外观特征和局部运动特征. 之后,将 TSN_G 网络提取的全局外观和全局光流特征进行特征融合;类似地,将 TSN_L 网络提取的局部外观和局部光流特征进行特征融合. 再则,两条支流 TSN 融合后的特征,分别通过各自的 LSTM 网络捕获长期依赖时序特征,并实现行为的预判断. 最后,将两路 LSTM 得到的行为识别结果进行分数融合,得到组群行为标签. 下面将分别详述其中的数据预处理、TSN 特征提取、特征融合及行为识别几个模块.

4 数据预处理

4.1 CAD 数据集预处理

组群行为识别的关键是捕获更高层级的特征表示. 在进行特征提取之前,必须对视频中的主要人物进行定位与跟踪等预处理操作. 即利用 CAD 中已提供的人体目标候选框,形成被跟踪人物的边界框序列,完成对视频中人物目标的跟踪. 如图 3 示例,为 CAD 中的三种行为剪辑帧,包括边界框、单人行为和组群行为的标定. 这些人体目标会作为局部特征,配合全帧图像的全局信息,实现本文提出的多视觉信息融合.

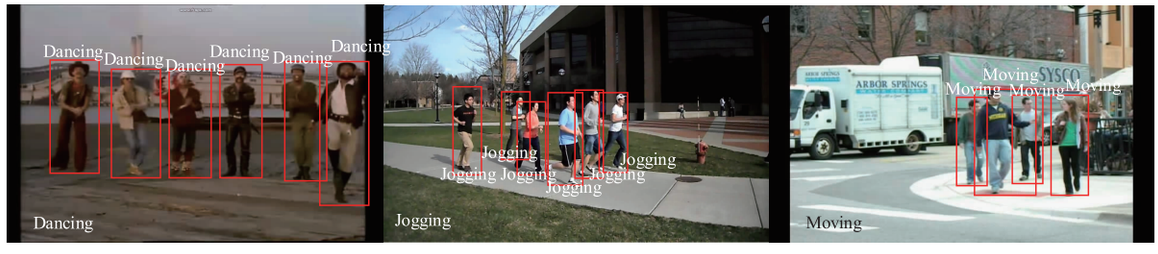

图3 CAD中的单人行为和组群行为的边界框标定示例

4.2 光流特征预处理

本文重点考虑多视觉线索输入,在光流特征提取方面,采用翘曲光流(Warped Optical Flow Fields)提取方法,提取视频中人物 X 方向和 Y 方向的光流信息,将捕获的运动特征作为 TSN 网络时间网的输入. X 和 Y 两个方向提取的光流用热度图表示如图 4 所示,其中最右边为热度标尺,将 Colorbar 归一化到[0,1]区间内,从 1 到 0 表示运动的强弱,由图可知人体目标候选框及整幅图片的 X 方向光流运动较强,Y 方向光流运动普遍较弱,这是因为当前场景中的人物是水平移动而不是垂直移动,因此 X 方向光流要比 Y 方向光流强度大.

图4 两种输入模态: 左边上下两图是局部人物候选框及其对应的X,Y两个方向的热度图;
中间上下两图是整张全局图片及其对应的X,Y两个方向的热度图; 最右边是热度标尺

5 TSN 特征提取

本文提出利用 TSN 网络对视频中多视觉线索的外观及光流信息进行特征提取. 该网络架构主要由空间网(Spatial ConvNets)和时间网(Temporal ConvNets)构成,采用稀疏采样的方法获取长视频中的短片段,样本沿时间维均匀分布,并采用分段结构从采样片段中聚合信息,使时间网能够对整个视频进行长时间建模,具体原理如下.

5.1 TSN 多视觉线索的特征提取与融合

假设给定一段视频 V,视频中共有 n 个人. 将它等间隔分为 K 段$\{S_1,S_2,\cdots,S_K\}$,本文使用的 TSN 网络原理公式如下:

$$\begin{aligned}&\mathrm{TSN}(T_1,T_2,\cdots,T_K)\\&=G(F(T_1;W),F(T_2;W),\cdots,F(T_K;W))\end{aligned}\tag{1}$$

其中,$(T_1,T_2,\cdots,T_K)$为片段序列,每一个子片段 T_K 都是从它对应的段 S_K 中随机采样得到,设 K 个子片段共有 M 帧;$F(T_K;W)$函数表示具有参数 W 的卷积网络,在短片段 T_i 上进行操作,提取每个子片段中随机采样的图像特征. 段共识函数 G 结合多个短片段的特征表示,即将子片段中提取的所有空间特征与运动特征连接起来. 并结合标准分类交叉熵损失,使用随机梯度下降(SGD)法对模型参数进行学习.

对于局部 TSN_L 来讲,提取到的 n 个人的时空特征记为$f_L=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$. 针对第 i 个人的 M 帧跟踪候选框图像,其输入到局部 TSN_L,可以得到稀疏空间特征 $\mathrm{person\_TSN}_S^i$,即 $\mathrm{person\_TSN}_S^i=[p_1^{i\_S},p_2^{i\_S},\cdots,p_M^{i\_S}]$;类似地,其 M 帧光流信息输入后可得到稀疏光流特征 $\mathrm{person\_TSN}_T^i$,即 $\mathrm{person\_TSN}_T^i=[p_1^{i\_T},p_2^{i\_T},\cdots,p_M^{i\_T}]$;这样第 i 个人时空特征级联融合为:

$$x_i=\mathrm{person\_TSN}_S^i\diamond\mathrm{person\_TSN}_T^i\tag{2}$$

该过程数据可视化如图 5 所示.

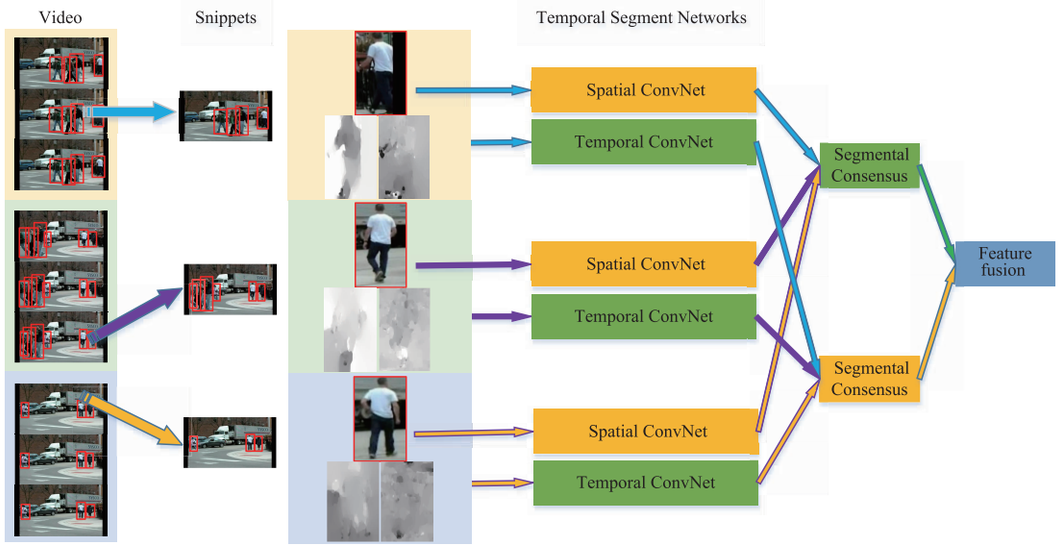

图5　以局部分支为例，第i个人的稀疏特征表示

则所有人时空特征矩阵表示为:

$$\begin{aligned}f_L&=x_1\diamond\cdots\diamond x_n\\&=\mathrm{person\_TSN}_S^1\diamond\mathrm{person\_TSN}_T^1\diamond\\&\cdots\diamond\mathrm{person\_TSN}_S^n\diamond\mathrm{person\_TSN}_T^n\end{aligned}\tag{3}$$

类似地,对于全局 TSN_G 来讲,M 帧场景 RGB 图像和场景光流图像的全局时空特征记为$f_G=\{t_1,t_2,\cdots,t_i,\cdots,t_M\}$,则第 i 帧的时空特征为:

$$t_i=\mathrm{scene\_TSN}_S^i\diamond\mathrm{scene\_TSN}_T^i\tag{4}$$

将 M 帧的场景时空特性使用级联融合策略,表示为:

$$\begin{aligned}f_G&=t_1\diamond\cdots\diamond t_M\\&=\mathrm{scene\_TSN}_S^1\diamond\mathrm{scene\_TSN}_T^1\diamond\\&\cdots\diamond\mathrm{scene\_TSN}_S^M\diamond\mathrm{scene\_TSN}_T^M\end{aligned}\tag{5}$$

6 组群行为上下文时序建模及行为识别

6.1 组群行为上下文时序建模

由于光流运动信息仅仅具有短期时序关系,因此,上面 TSN_G 和 TSN_L 网络提取的特征,不具有长时序关联性. 本文在双路 TSN 网络后分别连接 LSTM 网络,以期获得组群行为视频中的长时序上下文关系.

以局部特征分支为例,如图 6 所示,稀疏采样后在 $t-1$帧时,TSN_L 局部网络捕获人体目标的空间信息和光流运动特征,并级联融合;类似地,可得到第 t 帧和 $t+1$帧时的融合特征. 它们输入到 LSTM 网络中,捕获组群行为的长期时序依赖关系,生成用于组群行为识别的上下文综合特征描述.

该 TSN + LSTM 架构,前端捕获的光流微观运动信息,可以作为后续 LSTM 的补充;同时 LSTM 在后端提取宏观时序运动特征;相比较 CNN + LSTM 架构提取的时序特征更全面和细致.

6.2 组群行为识别

针对 CAD 数据集的特点,组群行为判别标准是将视频中大多数人的行为属性视为组群行为标签. 因此,本文采用决策融合方式,实现全局场景和个人局部信息的互补,进行组群行为识别.

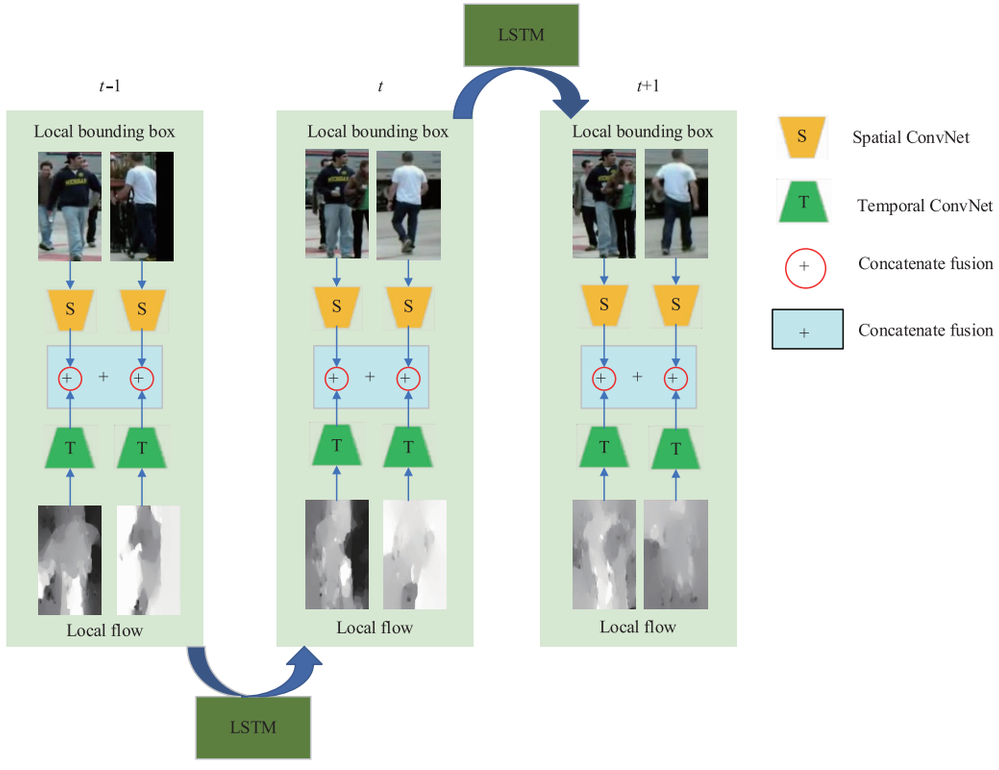

图6 以局部特征为例，上下文时序信息建模

将局部和全局支路经全连接层输出特征分别表示为 Model_G和 Model_L；两者经过 softmax 层实现行为识别，即

$$y_G = \text{softmax}(\text{Model}_G) \tag{6}$$

$$y_L = \text{softmax}(\text{Model}_L) \tag{7}$$

其中，y_G 表示全局支路行为识别的 softmax 得分，y_L 表示局部支路行为识别的 softmax 得分. 采用加权平均融合策略，获取最终组群行为标签 Y，即

$$Y = \text{AVG}(y_G, y_L) \tag{8}$$

在该策略下，全局特征与局部特征进行有效互补，实现全局特征与局部特征的融合对于消除个人行为的歧义具有重要作用，可进一步提高组群行为识别准确率.

7 算法验证

本文算法在 CAD1 和 CAD2 上进行了测试，下面 7.1 小节将简单介绍这两个数据集以及训练时的数据拆分安排；而实验内容设计、实验数据分析，以及与其它文献进行对比结果等将在 7.2 小节中详细展示.

7.1 数据集简介和拆分安排

7.1.1 数据集简介

CAD1[20]包含由低分辨率手持相机收集的 44 个视频，五类行为标签：Crossing，Waiting，Queuing，Walking，Talking；八种姿势标签（实验中未使用）；五种组群行为标签：Crossing，Waiting，Queuing，Walking，Talking. 根据大多数人在场景中所做的事情，场景被赋予小组行为的标签，以及每个人都有一个行为标签，每一帧图像都有一个场景行为标签. 在 CAD1 中，本文将“Walking”和“Crossing”合并为“Moving”，因此，本文 CAD1 中共有四类行为标签：Moving，Waiting，Queuing，Talking.

CAD2 是对 CAD1 的一个扩展，把 Walking 动作从 CAD 1 中移除，再补充 Dancing 和 Jogging 两个额外动作，这样本文 CAD2 共有六个动作类别，分别是 Crossing，Queuing，Dancing，Talking，Waiting，Jogging.

7.1.2 训练数据集的拆分安排

本文采用分段训练来学习模型参数. 即将 TSN 与 LSTM 分别独立进行训练，每一部分训练都是独立的. 并将上述 CAD1 和 CAD2 分别按照 3∶1 的比例进行划分，其中 3/4 的数据进行训练，剩余 1/4 的数据进行测试与验证，得出各类准确率以及平均识别精度（MPCA）.

7.2 实验内容安排和结果分析

7.2.1 实验环境及网络参数初始化

本文模型在 Caffe 框架下进行，ubuntu 版本为 14.04，cuda 版本为 8.0，cudnn 版本为 6.0，显卡型号为 NVIDIA GTX1080Ti. TSN 空间网以 RGB 图像作为输入，

图片尺寸设置为 224×224. 采用在 UCF101 split_1 上训练的预训练模型来初始化 RGB 和光流模型,防止网络过拟合,加速收敛. 并将空间流与时间流的比例设置为 1:1.5,空间网 K 值设置为 7,即把整段视频分为 7 段,每段随机挑选 1 帧;时间网 K 值设置为 3,即把整段视频分为 3 段,每段随机挑选连续 5 帧.

对于空间网来说,开始时,将学习率设置为 10^{-6},从 4000 次开始,每 2000 次迭代,将学习率减少到原来值的 1/10,总共迭代 9000 次. 对于时间网来说,将学习率设置为 5×10^{-6},从 5000 次开始,每 2000 次迭代,将学习率减少到原来值的 1/10,9000 次时停止下降,总共迭代 12000 次. 对于 LSTM 网络来讲,经过 TSN 网络后输出维度为 24×2048,每个 LSTM 层都包含 1024 个隐藏单元,并将 LSTM 的输出数量设置为类的数量.

7.2.2　CAD1 上各基线模型实验结果与分析

在实验阶段,本文设计了三个基线模型如下所示:

Baseline1 是在 TSN + LSTM 网络的基础上,只通过 TSN 网络提取全局特征,验证局部特征的有效性.

Baseline2 是在 TSN + LSTM 网络的基础上,只通过 TSN 网络提取局部特征,验证全局特征的有效性.

Baseline3 是在 TSN + LSTM 网络的基础上,只用 TSN 网络提取全局和局部特征,而不连接 LSTM 网络,验证 LSTM 网络长期时序信息的有效性.

最后是 TSN + LSTM 的完整网络模型,验证本文算法的有效性.

表 1 是 CAD1 上各种基线方法比较结果. 实验数据说明既提取全局特征又提取局部特征的方法优于基线方法的任意一种;同时也说明了 LSTM 网络具有捕获视频中长期依赖关系的能力.

表 1　模型在 CAD1 上各种基线方法的比较

Model	MPCA
B1-TSN(S_G+T_G) + LSTM	92.8%
B2-TSN(S_L+T_L) + LSTM	85.1%
B3-TSN($S_G+T_G+S_L+T_L$)	92.6%
TSN + LSTM (Ours)	93.2%

7.2.3　CAD1 上实验结果与其他方法的比较

表 2 将本文方法与目前较先进以及一些基本方法进行比较. 前四种方法都是没有将“Walking”和“Crossing”合并为“Moving”类的,由于“Walking”与“Crossing”类具有相似的视觉特征,因此误识别率比较大;而后三类方法是合并之后的,从而降低了这两类的误识别率.

表 2　模型在 CAD1 上的平均识别准确率(%)以及与其他方法的比较

Model	Crossing	Walking	Waiting	Queuing	Talking	MPCA
Two-stage Hierarchical Model[11]	61.54	80.41	66.44	96.77	99.45	81.5%
HANs + HCNs[21]	71	81	72	97	99	84.3%
Latent Variable Embedding[22]	88	33	88	98	99	85.4%
姿态特征 + 行为属性[23]	80	63	51	83	94	78.9%
Recurrent modeling[19]	94.39		63.64	100	99.45	89.4%
PC-TDM[12]	92.8		76.6	100	99.5	92.2%
Sructure Inference Machines[18]	–		–	–	–	81.2%
TSN + LSTM (Ours)	91.3		84.2	99	98.4	93.2%

文献[11]是利用双层 LSTM 网络分层模型,实现对组群行为的识别. 文献[12]提出一种参与贡献时间动态模型(PC-TDM),利用个人 LSTM 捕获个人动态信息,然后通过交互 Bi-LSTM 捕获人与人之间的交互信息,实现组群行为识别. 文献[21]是在文献[11]的基础上,对于部分/人级别特征提取上应用“分级注意网络”,该网络对于不同的人及其身体部位给予不同程度的关注度,并利用两层级 LSTM 网络对组群间的上下文关系进行建模,从而生成组群识别的高级特征表示,取得了不错的效果. 但本文 TSN + LSTM 模型在 CAD1 上的识别率均高于文献[11]、文献[12]和文献[21],主要是因为本文在捕获长期依赖的同时,更加注重多重视觉线索的输入,加入了人物的光流信息及全局特征表示,从而提高了识别的准确率. 本文 TSN + LSTM 模型在 CAD1 上的识别率也高于文献[19],文献[19]提出循环交互上下文建模方案,利用三层 LSTM 网络架构,对个人动态,组内和组间的交互关系进行统一建模. 文献[19]虽然在捕获个人动态信息时提取了个人空间信息和光流信息,但缺乏对全局特征的提取. 同时,本文模型明显高于传统手工设计特征[23]的方法,说明在人物特征提取和行为识别方面,本文模型具有优势.

7.2.4　CAD2 上实验结果与其他方法的比较

表 3 表明本文模型在 CAD2 上的平均识别率高于 CAD1 上的结果,一方面是因为 CAD2 中去除了 Walking 类,避免了 Crossing 误判成 Walking 的情况;另一方面是

由于新加入的 Dancing 和 Jogging 两个动作视频中几乎所有人都在干同一件事情,所以在 CAD2 上的识别率会高于在 CAD1 上的识别率.

表 3 在 CAD2 上的平均识别准确率(%)以及与其他方法的比较

Methods	MPCA
Structure Inference Machines[18]	90.23%
Latent Variable Embedding[22]	97.94%
姿态特征+行为属性[23]	83.1%
TSN + LSTM (Ours)	95.7%

与其他方法的对比,本文模型在 CAD2 上的识别精度高于文献[18],主要是因为虽然 Structure Inference Machines 方法考虑了人与人之间的交互关系,并引入可训练的门控函数来决定谁与谁互动,通过反复改进推测个人行为来聚合场景中其他人行为的线索,但文献[18]并没有将主要人物的运动信息及全局信息考虑进去. 本文模型准确率低于文献[22],主要是由于文献[22]中将潜在变量嵌入到特征空间,并在深度学习框架中利用特征映射函数,来参与集体场景的个体之间复杂结构依赖性的建模,从而达到了很好的识别效果,该方法适用性更广. 但在 CAD2 上,由于人与人之间的复杂依赖关系并不多,组群中的个体行为相对独立,因此本文模型的针对性更强;同时,本文模型依旧高于传统方法[23]. 因此在 CAD2 上同样表明了本文模型对组群行为识别的有效性和稳定性.

8 总结

针对组群行为识别问题,本文采用 TSN + LSTM 模型,利用两路 TSN 网络进行视频中局部和全局特征的提取,结合全局及局部的外观及运动信息,完成特征提取. 然后两路 TSN 网络特征融合后分别连接 LSTM 网络,生成用于组群行为识别的具有上下文时序信息的高级特征表示,最终实现组群行为识别.

此外,在 CAD1 和 CAD2 上对本文模型进行训练和测试,并与多种不同的方法进行对比分析,证明了光流信息对于人物特征提取的重要性,同时也证明了全局特征与局部特征相结合,可以进一步提升行为识别的准确性,从而表明该算法对组群行为识别的有效性和稳定性. 未来,我们计划优化网络模型,使得该算法在更多的组群行为相关数据集上得到更优的实验结果.

参考文献

[1] Carreira J, Zisserman A. Quo Vadis, action recognition? A new model and the kinetics dataset[A]. Proceedings of CVPR[C]. USA: IEEE, 2017. 6299 – 6308.

[2] Sudhakaran S, Escalera S, Lanz O. LSTA: Long short-term attention for egocentric action recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2019. 9954 – 9963.

[3] YAN An, WANG Yali, LI Zhifeng, QIAO Yu. PA3D: Pose-Action 3D machine for video recognition[A]. Proceedings of CVPR[C]. USA: IEEE, 2019. 7922 – 7931.

[4] 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.

[5] Yuan Y, Wang D, Wang Q. Memory-augmented temporal dynamic learning for action recognition[A]. The Thirty-Third AAAI Conference on Artificial Intelligence[C]. AAAI, 2019. 9167 – 9175.

[6] 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 Guo-dong. Gaussian mixture models and principal component analysis based human trajectory behavior recognition[J]. Acta Electronica Sinica, 2016, 44(1): 143 – 149. (in Chinese)

[7] DING Wenwen, LIU Kai, XU Biao, CHENG Fei. Skeleton-based human action recognition via screw matrices[J]. Chinese Journal of Electronics, 2017, 26(4): 790 – 796.

[8] 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)

[9] 郑兴华,孙喜庆,吕嘉欣,鲜征征,李磊. 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, XIAN Zheng-zheng, LI Lei. Action recognition based on deep learning and artificial intelligence planning[J]. Acta Electronica Sinica, 2019, 47(8): 1661 – 1668. (in Chinese)

[10] 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.

[11] 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.

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

[13] Simonyan K, Zisserman A. Two-stream convolutional net-

works for action recognition in videos[A]. NIPS 2014[C]. NIPS,2014. 568 – 576.
[14] Wang L,Xiong Y,Wang Z,et al. Temporal segment networks:towards good practices for deep action recognition[A]. European Conference on Computer Vision[C]. ECCV,2016. 20 – 36.
[15] Zhu Y,Lan Z,Newsam S,et al. Hidden two-stream convolutional networks for action recognition[A]. Asian Conference on Computer Vision[C]. ACCV, 2018. 363 – 378.
[16] Ramanathan V,Huang J,Abu-El-Haija S,et al. Detecting events and key actors in multi-person videos[A]. Proceedings of CVPR[C]. USA:IEEE,2016. 3043 – 3053.
[17] Qi M,Qin J,Li A,et al. stagNet:an attentive semantic rnn for group activity recognition[A]. European Conference on Computer Vision[C]. ECCV,2018. 101 – 118.
[18] 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.
[19] Wang M,Ni B,Yang X. Recurrent modeling of interaction context for collective activity recognition[A]. Proceedings of CVPR[C]. USA:IEEE,2017. 3048 – 3056.
[20] 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 Workshops,ICCV Workshops[C]. USA:IEEE,2009. 1282 – 1289.
[21] 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.
[22] Tang Y,Zhang P,Hu J F,et al. Latent embeddings for collective activity recognition[A]. Proceedings of the 14th IEEE International Conference on Advanced Video and Signal based Surveillance[C]. USA:IEEE,2017. 1 – 6.
[23] 豆贺贺. 基于视频特征的多人行为识别研究[D]. 南京:南京邮电大学,2016. 1 – 61.

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

胡小悦　女,1993 年 8 月出生,山东济南人. 2016 年毕业于青岛工学院信息学院,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:1783910733@qq.com

孟唯佳　女,1995 年 4 月出生,山东济南人. 2017 年毕业于山东师范大学应用化学系,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:386345026@qq.com

闫春娟　女,1969 年 1 月出生,山东莱阳人. 1991 年毕业于中国石油大学物理勘探系,取得学士学位,2000 年至今在青岛科技大学信息学院通信教研室工作,从事信息与通信系统方面的有关研究.
E-mail:qdyancj@163.com