ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 4,Jul. 2019,pp. 585－593
DOI:10. 16337/j. 1004-9037. 2019. 04. 002

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于ActionVLAD池化与分层深度学习网络的组群行为识别方法

王传旭　姜成恒

(青岛科技大学信息科学技术学院,青岛,266100)

摘　要：构建端到端的深度学习网络结合局部聚合描述符(Action vector of locally aggregated descriptor, ActionVLAD)池化层和多层长短时记忆(Long short time memory,LSTM)解决组群行为识别问题。在传统的单一图像信息(Red Green Blue, RGB)作为深度学习网络的输入基础上,添加密集光流信息(Dense_flow),描述视频帧间的运动,作为双流网络的输入;通过底层LSTM对特征信息进行建模,由融合的双流特征来表示个人行为;而ActionVLAD池化层可以对不同时间、图片不同位置的特征进行融合,从而更好地融合个人信息;最后顶层LSTM连接Softmax分类器,通过融合的个人信息判断组群活动。在Collective activity dataset数据集上的测试实验获得了82.3%的平均识别精度。

关键词：组群行为识别;局部聚合描述符;双流网络;分层长短时记忆

中图分类号：TP391　　**文献标志码**:A

Group Behavior Recognition Method Based on ActionVLAD Pooling and Hierarchical Deep Learning Network

Wang Chuanxu, Jiang Chengheng

(School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, 266100, China)

Abstract: In group behavior recognition, the entire group behavior can be inferred by detecting the behavior of each person in the group over a period of time. An end-to-end deep learning network combined with action vector of locally aggregated descriptor (ActionVLAD) pooling layer and multi-layer long short time memory (LSTM) is constructed to solve the group behavior recognition problem. Based on the input of traditional single image information (Red Green Blue, RGB) as a deep learning network, dense optical flow information (Dense_flow) is added to describe the motion between video frames as the input of the two-stream network. The feature information is modeled by the underlying LSTM, and the individual behavior is represented by the fused two stream features. While the ActionVLAD pooling layer can fuse features at different time and different positions of the picture, which can better integrate personal information. Finally the top LSTM is connected with the Softmax classifier, in which group activity is judged by the merged personal information. The test on Collective activity dataset obtains an average recognition accuracy of 82.3%.

Key words: group behavior recognition; actionVLAD; two-stream network; hierarchical LSTM

基金项目:国家自然科学基金(61672305)资助项目;国家青年科学基金(61702295)资助项目。
收稿日期:2018-12-11;**修订日期**:2019-04-15

引 言

人体行为识别是计算机视觉领域一个重要的研究方向。随着卷积神经网络(Convolutional neural networks, CNNs)[1-3]的出现和日益成熟,物体识别与分类已经达到很高水平,相比之下,在人的行为识别中,特别是组群行为识别,识别精度不高。动态数据集不足,组群行为中多人的跟踪和识别问题,人与人之间的复杂关系建模等都是巨大的挑战。因此,组群行为识别受到研究者的广泛关注。

组群行为中最简单的识别方法是将视频帧整体作为深度学习网络的输入,根据发生的组群活动来训练模型,达到组群识别的目的。但是,直接训练整体图像,目标干扰物太多,并不能得到很好效果。目前用作组群行为识别最常见且有效的网络框架有两类:一类是直接将视频作为输入的3D卷积网络(3D spatio-temporal convolutions)[4],可以学习比较复杂的时空依赖关系,但性能不好,难以扩展;另一类即是将图像以及光流分别作为输入的双流网络,将视频帧分为外观流和运动流,分别训练其对应的CNN,最后融合输出的特征信息。目前双流网络的方法在效果上还是明显要优于3D卷积网络的方法。

本文采用双流网络的方法构建了一个端到端的分层深度学习网络模型。首先通过双流网络对组群中的每个人进行建模,然后利用局部聚合描述符(Action vector of locally aggregated descriptor, ActionVLAD)池化层对提取到的个人特征进行融合,根据组群中个人的融合特征表示团体活动。

近几年,深度学习网络广泛用于组群行为识别研究,大大解决了传统方法对复杂人群的限制和识别精度低的问题。Lan等[5]提出了一种自适应潜在结构学习识别组群活动,它能捕获组群活动及个体行为和他们之间的相互作用。Choi等[6]提出统一跟踪多个人,在一个联合框架中识别个人行为、互动和集体活动。Ibrahim等[7]使用分层深度时间模型来聚合用于整个活动理解的人员级信息。Wang等[8]利用长短时记忆(Long short time memory,LSTM)统一了单人动态,组内和组间交互的交互特征建模过程。Yao等[9]提出了一个多粒度交互预测网络,它集成了全局运动和详细的局部动作。以上所有方法关注的是个体建模,通过交互和集体活动来识别组群行为,虽然很大程度提高了组群识别精度,但是网络输入单一(RGB图像),建模过于复杂,无法实现端到端的训练。借鉴以往方法,本文提出端到端的分层深度学习网络框架,利用双流网络连接底层LSTM提取并表示组群中个人特征,结合ActionVLAD池化层融合个人级信息,经过顶层LSTM连接分类器实现组群活动识别。

1 算法描述

1.1 模型概述

本文算法模型流程如图1所示。本文的目标是检测识别多人在视频序列中的集体行为。组群活动识别问题的许多经典方法是基于人工设计的特征以结构化预测的形式对组群活动进行建模,受深度学习的启发,本文提出了多层次双流的深度学习网络。网络的输入为场景中每个人跟踪的RGB图像和Dense_flow图像(包括x和y分量),i表示组群中目标人数,用R_i表示第i个人的RGB图像,F_{ix}和F_{iy}表示第i个人的Dense_flow的x和y方向图像,本文使用Alexnet[10]和LSTM网络搭建双流网络对数据进行训练和学习,提取的双流特征由前向特征融合[11]算法模拟个人活动的时间动态表示,并获得个人行为的初步预测;再经过ActionVLAD池化层进行聚类得到组群行为的预估计,通过顶层LSTM层,保证特征信息不丢失的同时加强组群图像帧间的联系;最后,输出连接到Softmax分类层来检测视频序列中的组群活动类。本文创新之处是利用ActionVLAD池化聚类算法代替传统的池化算法对个人级行为特征进行聚类融合得到组群级行为的表示;利用分层的LSTM网络实现对个人级及组群级行为的建模表示,进而实现组群行为识别。

1.2 个人级双流特征融合

针对组群中的每个人的位置,对应提取其边界框信息,得到该组群行为的每个人的行为视频帧,作

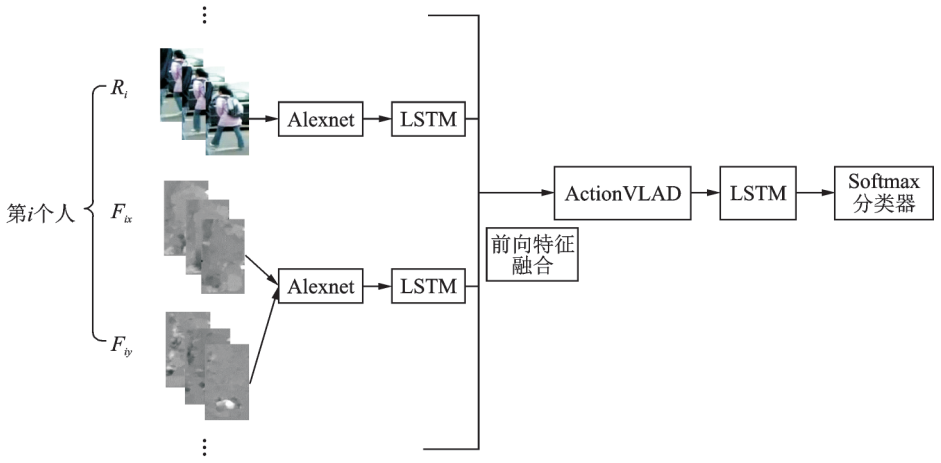

图1 多层双流的端到端网络流程图

Fig.1 Flow chart of multi-layer two-stream end-to-end network

为深度学习网络的输入。在行为识别中,不同的特征对应到相应的行为,有的特征得到的是它的静态特征比,如图像RGB特征,而密集光流特征可以得到人体行为的运动特征。不同的特征描述的是行为的不同方面,研究实验表明,把不同的特征融合起来可以得到更好的结果。

1.2.1 双流深度学习网络的搭建

基于传统的Two-stream CNN网络,利用CNN网络搭建双流网络,用作提取个人特征。如图2所示,应网络输入的需求,提取数据集对应的光流视频帧图像,将原数据集和光流数据集作为双流网络的

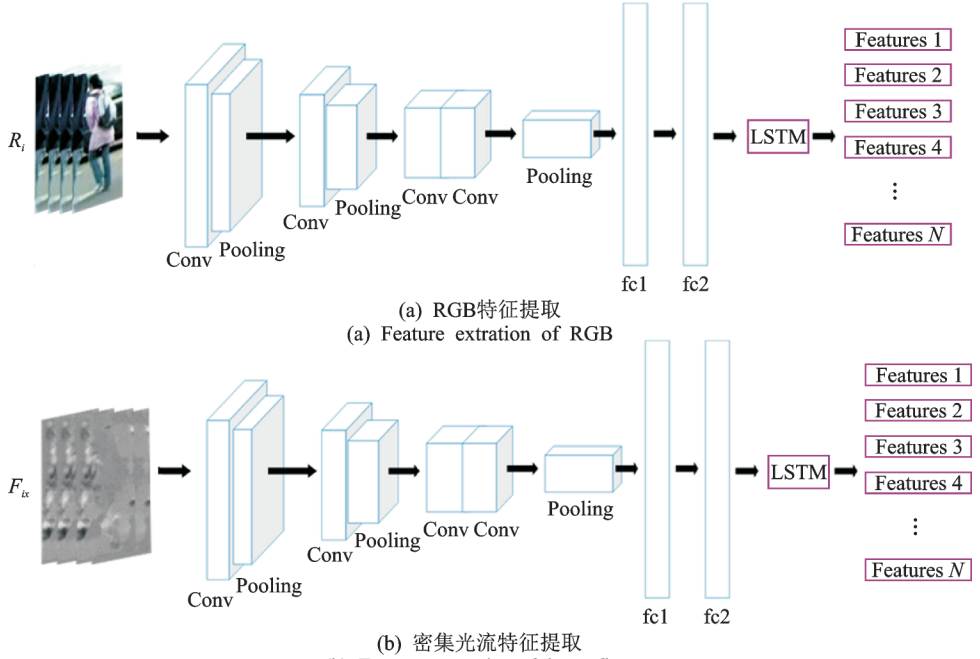

图2 双流网络特征提取

Fig.2 Feature extraction of two-stream network

输入，与传统双流网络不同的是，本文在经过卷积池化最后分别连接两层全连接层和一层LSTM层，充分保证提取特征的完整性以及添加时间联系，保证时间信息不丢失，最后通过一定的融合方法将双流网络提取的不同特征进行融合汇总。

1.2.2 双流特征融合

由于双流网络是独立的两个网络，得到的特征也是相对独立的，为了实现双流网络提取特征的丰富性，需要对特征进行进一步地融合。Ballan[11]提出了两种融合方式：一个是前向融合，就是在描述符的水平上进行融合，简单来说就是把两种不同的特征以串联的形式连接起来形成新的描述符；另外一种是后向融合，在对每种特征建立描述符之后，把两种不同特征得到的不同的行为描述符连接起来形成新的描述符。根据前人经验和本文需求，选用前向融合方式，如图3所示。

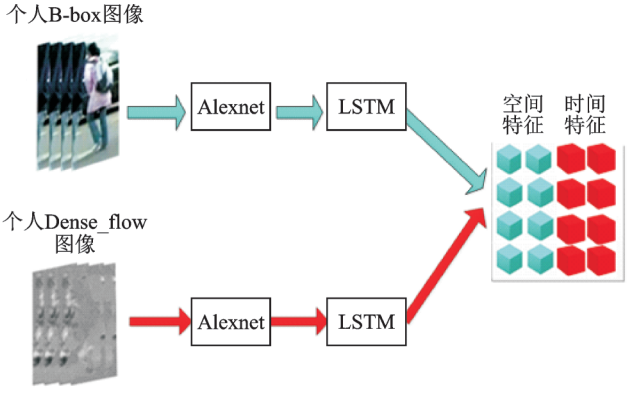

图3 个人前向特征融合

Fig.3 Personal forward feature fusion

1.3 ActionVLAD池化聚类的组群行为优化识别

通过本文搭建的双流网络得到了组群场景中的每个人的双流特征，通过前向特征融合，进而得到个人级特征描述和个人行为的预测，但是，仅仅描述和预测个人的行为并不能很好地达到识别组群行为的目的。针对这一问题，本文提出了使用ActionVLAD池化算法来聚合该组群行为中每个人的行为，通过对每个人建模分组，相同行为特征的个人被分到一个团体，由该场景中最大组群的行为来定义组群行为。

1.3.1 ActionVLAD池化聚类

Relja等[12]从局部聚合描述子向量(Vector of locally aggregated discriptor, VLAD)[13]在图像检索中良好的鲁棒性受到启发，提出了在CNN框架中模拟VLAD并设计可训练的广义的VLAD层，即NetVLAD时空聚合层。在NetVLAD聚合的时空扩展基础上，引入了时间t之和，称为ActionVLAD。假设$x_{it}\in\mathbf{R}^D$，是从一段视频帧$t\in\{1,\cdots,T\}$的位置$i\in\{1,\cdots,N\}$中提取一个D维局部描述符。将特征描述子空间$\mathbf{R}^D$划分为K个“动作词”区域，用锚点$\{c_k\}$表示，那么每一个描述子x_{it}被分配到一个聚类中心并由残差向量$x_{it}-c_k$表示，然后将所有残差向量累加用作表示整个视频，即有

$$V[j,k]=\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{e^{-\alpha\left\|x_{it}-c_k\right\|^2}}{\sum_{k'}e^{-\alpha\left\|x_{it}-c_{k'}\right\|^2}}(x_{it}[j]-c_k[j]) \tag{1}$$

式中：$x_{it}[j]$和$c_k[j]$分别为描述符向量$x_{i,t}$和锚点c_k的第j个分量；α是一个可调超参数；$c_{k'}$表示理论锚点数；k'表示理论“动作词”数量；上标T表示视频总帧数；N表示视频中总人数。式中第1项表示描述符

x_{it}到单元K的软分配,第2项$x_{it}[j]-c_k[j]$表示描述符和单元K的锚点之间的残差,两个求和运算符分别表示时间和空间的聚合,输出是矩阵V,表示k个聚类中心的D维特征描述子,经过归一化后展开为$v\in \mathbf{R}^{KD}$描述子即可表示整个视频,如图4所示。

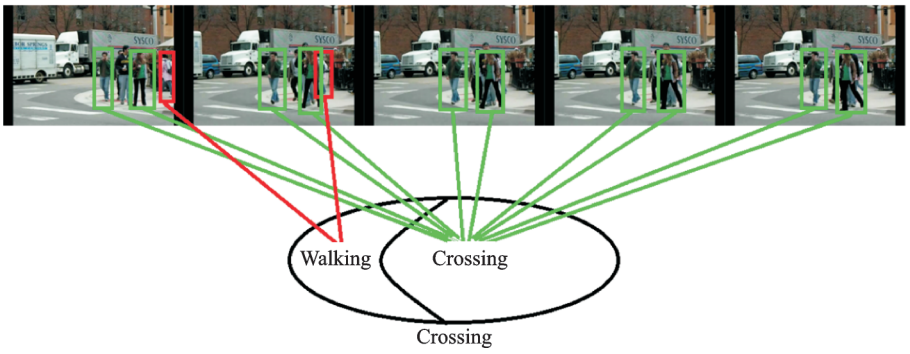

图4 ActionVLAD池化聚类

Fig.4 ActionVLAD pooling clustering

图4将目标人物特征根据ActionVLAD池化聚类为行走(Walking)和(Crossing)两类,依据视频中目标人物呈现的大部分状态定义组群行为为过马路(Crossing)。

从图5不同的池化策略建模对比可以看出,最大池化和平均池化都只能关注到部分子类特征,而ActionVLAD却可以聚合不同子类特征的描述子来共同描述视频特征。在组群活动中,个体一般呈现不同的行为。例如,本文数据集标签为过马路(Crossing)的场景中,有不少行人的状态是行走(Walking),通过ActionVLAD池化,可以更好地区分组群活动中小团体活动,从而更准确定位组群活动。

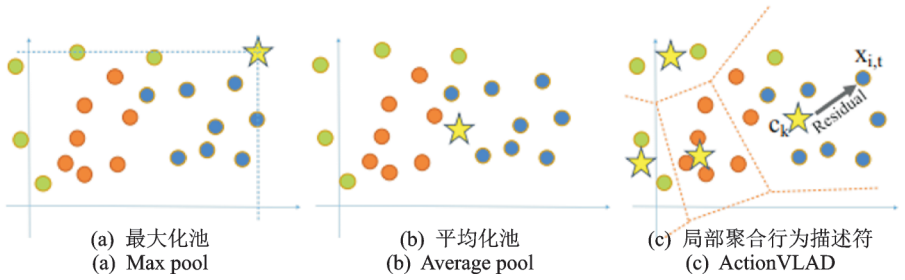

图5 不同池化策略建模对比

Fig.5 Comparison of different pooling strategies

1.3.2 分层LSTM对个人及组群行为的优化

本文构建的双流CNN网络,不仅可以提取个人动作特征(密集光流特征),同时提取到的基于个人图像周围空间域的特征(RGB特征),也作为动作的判别信号之一。特征提取后,在每个时间步骤中,CNN网络连接底层LSTM层用于提取个人的动作信息以及其动作中的时间变化的信息,其得到的输出特征向量表示个人的时间活动特征。把同一组群场景中目标人定义为$x=\{x_1,x_2,\cdots,x_N\}$,其中x_i表示第i个人的特征信息,包括本文利用双流深度学习网络提到的外观特征信息和动作信息;N为视频中该组群中个人的数量;定义相应的行为分类为$y=\{y_1,y_2,\cdots,y_N\}$,每个变量y_i所对应的标签集合为$L=\{l_1,\cdots,l_K\}$,K为行为分类的数量,这里可以简单地认为$x=\{x_1,x_2,\cdots,x_N\}$,即为底层LSTM网络对个人建模得到的特征向量,而后经过Softmax分类层,实现对个人行为的初步预测。

使用ActionVLAD池化层聚合算法对时间T中场景里所有人的特征进行聚类汇总。如式(1)所示，v表示该时间场景下组群行为的特征描述，应数据集分类要求，本文在训练验证过程中，将聚类最多的分类特征作为组群特征描述，即$v=\max\{f(V[\cdot,k_0],\cdots,V[\cdot,k_n]),n\in K-1\}$，$K$为行为分类的数量，$x=f(x_1,\cdots,x_n)$，即求特征最多的矩阵，假设$n=2$，将该组群划分为3个类别团体，同时$n=1$时聚类特征最多，则组群描述符$v=V[\cdot,k_1]$表示该组群行为。最后池化层的输出作为顶层LSTM的输入，聚合得到的特征向量用作组群行为的表示，判断依据为场景中绝大多数人的行为即为组群的行为。在时间表示之上工作的顶层LSTM网络直接连接到分类层，用于直接模拟组群活动的时间动态，以便检测视频序列中的组活动类，如图6所示。

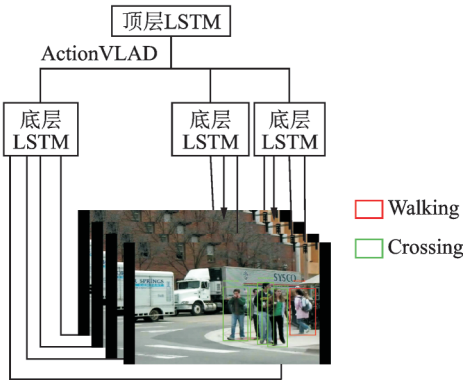

图6 分层LSTM建模框架

Fig.6 Hierarchical LSTM modeling framework

2 算法验证

基于Collective Activity Dataset1数据集，实验硬件及软件环境为：深度学习服务器使用Intel Core i7-5960X(主频3.0GHZ)；图形加速使用Nvidia GPU(型号GeForce GTX 1080 8G)；Linux(Ubuntu 14.04)操作系统和Linux版本的MATLAB 2015b；编程语言包括Python脚本语言以及MATLAB编程语言。

为了验证本文模型的可行性，实验内容如下：(1)简要介绍CAD数据集；(2)阐述实验中网络参数的配置与选择；(3)为验证模型中ActionVLAD池化算法以及双层LSTM的作用，设计了与3组Baseline方法的对比，其中Baseline1是在本文搭建的双流深度学习网络和双层LSTM网络基础上，利用传统的平均池化算法替代本文的ActionVLAD算法，Baseline2是在本文搭建的双流网络和底层LSTM网络基础上，剔除顶层的LSTM，Baseline3是在双流网络和顶层LSTM网络基础上，剔除底层的LSTM；(4)最后将本文模型与state-of-the-art方法进行对比，同时进行损失函数和精度的性能分析。

2.1 组群行为数据集

实验采用组群行为识别数据集(Collective activity dataset，CAD)。CAD是使用低分辨率的手持摄像机获取的44个视频片段，此数据集有5种行为(Action)标签：Crossing，Queuing，Walking，Talking和Waiting；8种姿势标签(实验中未使用)；5种活动标签即每帧活动中N个人共同完成的场景标签：Crossing，Queuing，Walking，Talking和Waiting。每个人都有1个行为标签，每帧图像都有1个场景活动标签，如图7所示。

实验中，对于上述两个数据集都是与文献[14]的分裂方式，其中1/3用于测试，其余用于验证与测试。数据集中的K个Action需要进行K组测试与计算，然后对得到的这K个测试结果求平均值，得到最终的平均识别精度。

2.2 网络参数的配置与选择

2.2.1 网络的配置

训练本文的模型大致分为2步：(1)使用由动作标签注释的个人轨迹组成的RGB图像和Dense_flow图像作为训练数据以端到端的方式训练个人级CNN和第1 LSTM层。本文使用Caffe实现框架模型，借鉴文献[7]设定网络参数，第1个LSTM层使用9个时间步长和3 000个隐藏节点，使用预先训练的AlexNet网络初始化CNN模型，并且为第1个LSTM层微调整个网络，提取到的双流特征经过前向特征融合形成个人级特征描述，并通过Softmax层进行分类识别，得到个人级的行为预测；(2)对场

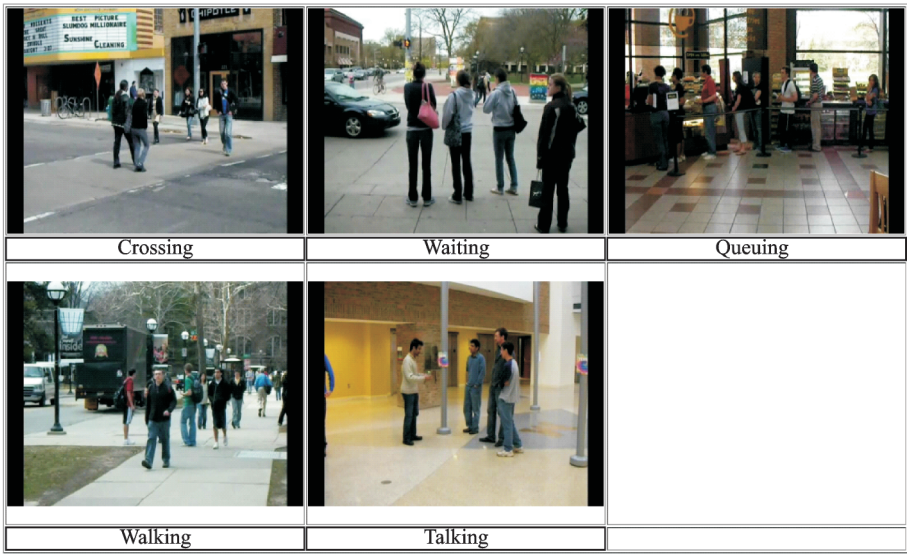

图7 CAD数据集

Fig.7 CAD dataset

景中的所有人特征进行ActionVLAD池化聚类,送到第2层LSTM网络,该网络由1个3 000节点的全连接层和9个时间步长的500个隐藏节点的节点LSTM层组成,用于添加组群帧间的联系,最后将其传递给识别组群活动标签的Softmax层,得到最终的分类结果。

2.2.2 参数的选择

在算法模型的参数训练过程中,初始学习率设定为0.000 1,每迭代20 000次减小0.1倍,最大迭代次数为500 000次,冲量(Momentum)设为0.9。因为在使用Gaussian初始化方法对网络进行训练时,实验中由于两个数据集在深度学习网络中参数很难收敛,最终的损失(Loss)值非常大,因此,本文在卷积层和全连接层的初始化方式上分别选用MSRA和Xavier,替代了原始CNN中所使用的Gaussian初始化方式。第一阶段的训练,本文采用Finetune的方式,在已经训练好的模型参数基础上,然后在所使用的数据集上进行进一步的参数优化,这样可以节省训练时间,使得网络可以尽快地收敛,获得最优的实验结果。

2.3 实验结果及与其他算法的对比

2.3.1 本文模型实验结果分析

本文模型在CAD数据集获得的混淆矩阵如图8所示,观察到本文模型实现了对组群行为识别,几乎完美地用于说话和排队类,但是在Crossing,Waiting和Walking之间混淆。这种情况可能是由于缺乏对该组中人们之间的空间关系的考虑,例如,Crossing类中涉及Walking类,但与Walking类最大的不同是人们以有序的方式在规定的道路中行进。而本文模型只是为了学习组群活动的动态属性,因此导致两类动作混淆。

	Crossing	Waiting	Quening	Walking	Talking
Crossing	65.34	4.13	0.74	29.79	0.00
Waiting	11.41	67.44	0.00	21.15	0.00
Queuing	0.00	0.00	96.77	3.23	0.00
Walking	15.49	2.09	0.00	82.41	0.00
Talking	0.00	0.00	0.00	0.45	99.55

图8 本文模型获得的CAD数据集的混淆矩阵

Fig.8 Confusion matrix for CAD dataset obtained by using the proposed model

2.3.2 与Baseline和其他算法的识别精度对比

本文模型(Ours)在CAD数据集中各类活动的识别

率高低以及与state-of-the-art和3组Baseline方法对比的结果如表1所示。表1中展示了多个组群行为识别模型在CAD数据集的平均识别准确率,通过比较,可以看出本文模型的平均识别准确率高于其他模型。CAD数据集包括组群活动Crossing,Queuing,Walking,Talking和Waiting,由于Waiting类的组群定义不明确,更偏向于单人行为识别,而不是组群行为识别,这是导致Waiting类识别率不高的主要原因;Walking类和Crossing类之间的唯一区别就是人与街道之间的关系,这是导致两者识别精度较低的最主要因素。

表1 模型在CAD数据集上的平均识别准确率以及与其他方法比较

Tab. 1 Average recognition accuracy of different models on CAD dataset %

Class/Model	文献[15]	文献[16]	Baseline1	Baseline2	Baseline3	Ours
Crossing	61	88	59.53	61.48	63.82	65.34
Waiting	66	88	66.21	64.95	65.98	67.44
Queuing	96	98	87.84	92.36	95.23	96.77
Walking	80	33	55.76	74.03	84.79	82.41
Talking	99	99	93.67	94.18	97.68	99.55
平均精度	80.9	81.2	72.6	77.4	81.5	82.3

在Waiting类中,本文模型与基线方法的识别率明显低于文献[15]方法,究其原因,本文没有使用数据集自带的单人姿势标签信息。此外,CAD数据集采集来源与真实的社会场景,Waiting类总是伴随着Crossing类和Walking类同时出现,也是导致混淆预测的一个重要的因素。使用双流深度学习网络方法时序模型的识别结果高于传统的基于RGB图像特征的深度学习网络[16]的方法,说明在特征提取过程中,多特征的融合在一定程度上可以提高识别精度;在个人级建模上升到组群级建模过程中,ActionVLAD池化聚类方法优于最大池化方法。本文设定的Baseline1,Baseline2和Baseline3方法的识别精度低于本文模型,通过3种基线方法与本文模型的比较,可以得到ActionVLAD池化,很好地解决了个人到组群中聚类的问题;顶层LSTM(组群级LSTM)较底层LSTM(个人级LSTM)优先级更高,同时验证了每一层LSTM对个人及组群行为建模的必要性。

2.3.3 损失函数性能分析

CAD数据集在训练和测试时的损失(Loss)和精度(Accuracy)曲线如图9所示。由图中可以看出,CAD数据集在本文模型训练过程中迭代5万次的时候损失达到最低,且趋于稳定;测试精度也在5万以后接近稳定。

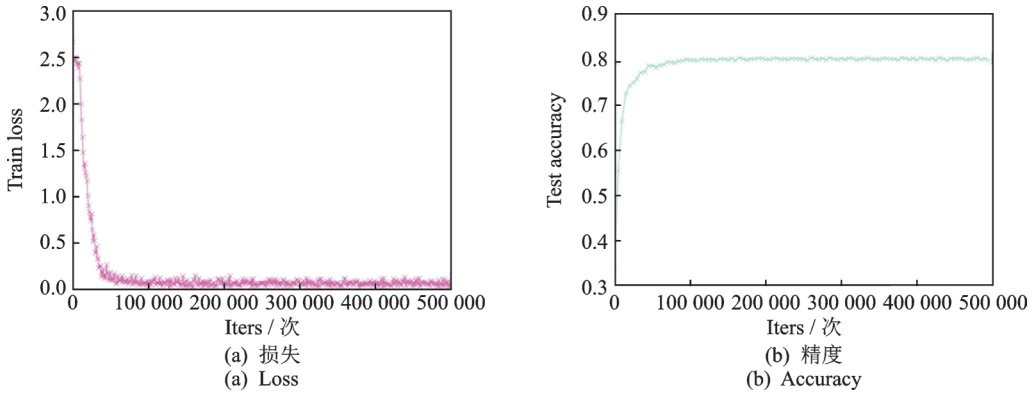

图9 CAD数据集的训练和测试曲线

Fig.9 Training and test curves of CAD dataset

3 结束语

本文提出了一种多阶段深层网络结构来处理组群活动识别问题。模型优势是通过两个阶段的过程学习了个人层级上的时间表示,利用ActionVLAD池化方法结合个体的表征来识别组群活动,并在CAD数据集上得到了较好的识别率;但是模型缺乏组群中人与人之间的空间联系,导致部分分类结果混淆较为严重。在下一步研究中,将对CAD数据集中的个人位置和姿态信息进行分析,加强人与人之间的联系,从而提高识别率。

参考文献:

[1] Girdhar R, Ramanan D, Gupta A, et al. ActionVLAD: Learning spatio-temporal aggregation for action classification[C]//Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA:IEEE, 2017: 3165-3174.

[2] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l]:IEEE, 2015: 4305-4314.

[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA: IEEE, 2016: 770-778.

[4] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. [S.l.]: IEEE, 2015: 4489-4497.

[5] 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, 2012, 34(8): 1549-1562.

[6] Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition[C]//European Conference on Computer Vision. Berlin, Heidelberg: Springer, 2012: 215-230.

[7] Ibrahim M S, Muralidharan S, Deng Z, et al. A hierarchical deep temporal model for group activity recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2016: 1971-1980.

[8] Wang M, Ni B, Yang X. Recurrent modeling of interaction context for collective activity recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 3048-3056.

[9] Yao Taiping, Wang Minsi, Ni Bingbing, et al. Multiple granularity group interaction prediction[C]//The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.]: IEEE, 2018:2246-2254.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc, 2012:1097-1105.

[11] Ballan L, Bertini M, Bimbo A D, et al. Effective codebooks for human action representation and classification in unconstrained videos[J]. IEEE Transactions on Multimedia, 2012, 14(4): 1234-1245.

[12] Arandjelovic R, Gronat P, Torii A, et al. NetVLAD: CNN architecture for weakly supervised place recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l]:IEEE, 2016: 5297-5307.

[13] Jégou H, Douze M, Schmid C, et al. Aggregating local descriptors into a compact image representation[C]//CVPR 2010-23rd IEEE Conference on Computer Vision &. Pattern Recognition. [S.l]:IEEE Computer Society, 2010: 3304-3311.

[14] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials[C]//Advances in Neural Information Processing Systems. [S.l]:[s.n.], 2011: 109-117.

[15] Tang Y, Zhang P, Hu J F, et al. Latent embeddings for collective activity recognition[C]//2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). [S.l.]: IEEE, 2017: 1-6.

[16] Ibrahim M S, Muralidharan S, Deng Z, et al. A hierarchical deep temporal model for group activity recognition[C]//Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on. [S.l.]: IEEE, 2016: 1971-1980.

作者简介:

王传旭(1968-),男,教授,研究方向:计算机视觉,E-mail:Wangchuanxu_qd@163.com。

姜成恒(1993-),男,硕士研究生,研究方向:计算机视觉。

(编辑:刘彦东)