文章编号：2096-1618(2022)01-0040-06

基于 ResNeXt-GRU 和聚类采样的人体行为识别

曾庆喜， 彭 辉
(成都信息工程大学软件工程学院，四川 成都 610225)

摘要：为有效捕捉行为中的时序关系，增强网络的特征表达能力，提出一种基于 ResNeXt-GRU 的人体行为识别方法。首先，使用聚类算法提取行为视频关键帧序列，输入 ResNeXt 网络中进行空间维度上的特征提取。然后，将输出的特征向量全部输入门控循环单元(GRU)网络中进行时序学习。最后，利用 Softmax 分类器进行分类。在 UCF101 和 HMDB51 数据集上分别进行实验，识别准确率为 93.7% 和 69.2%。实验结果表明与现有的其他许多行为识别方法相比，识别准确率得到了一定的提升。

关 键 词：行为识别；聚类；ResNeXt；门控循环单元(GRU)；Softmax
中图分类号：TP391 **文献标志码**：A
doi：10.16836/j.cnki.jcuit.2022.01.007

0 引言

近年来，随着计算机视觉技术的不断发展，人体行为识别已成为研究热点，在视频分类、视频监控、无人驾驶、人机交互等领域都具有广阔的应用前景[1]。人体行为是发生在特定时空的事件，行为特征不仅具有空间性，也具有时间性，如何有效地描述时空特征是行为识别问题的关键。基于卷积神经网络(CNN)在图像特征提取和循环神经网络(RNN)在处理时序问题有突出成果，研究者对行为识别任务提出了很多研究思路和方法。

Simonyan K 等[2]利用人体行为具有时空信息的特点，设计了一种时空双流卷积神经网络，在空间流和时间流上使用单独的二维卷积神经网络提取特征，最后通过 SVM 分类器进行分类。Wang 等[3]通过对视频分段和稀疏采样，提出一种时间片段 TSN 网络。Tran D 等[4]提出 C3D 网络，使用三维卷积和三维池化直接处理输入的人体行为视频，该模型耗费的时空资源较多，训练难度较大。Donahue J 等[5]通过结合卷积神经网络和循环神经网络变体长短时记忆模型(LSTM)提出了长时循环卷积神经网络(LRCN)。该方法先通过 CNN 提取行为特征，再通过一个 LSTM 网络提取时序信息，最后，通过 softmax 分类，LRCN 模型识别准确率与双流网络相比较低。Zhao 等[6]将改进并结合注意力机制的 CNN 和 RNN 相结合解决动作识别任务，也取得了不错的效果。

基于以上分析，在 LRCN 模型基础上，提出一种基于 ResNeXt-GRU 和聚类采样的人体行为识别方法。采用聚类算法思想改进视频帧采样方式，减少冗余数据输入，提升方法效率。使用 ResNeXt 深度卷积神经网络，加强人体行为空间特征提取的同时防止网络退化。结合 GRU 网络，进一步提取行为的时序性特征。利用 Softmax 分类器对人体行为进行分类。

收稿日期：2021-08-29
基金项目：四川省科技计划资助项目(2019YJ0356)

1 方法

1.1 概述

设计的人体行为识别方法整体流程如图 1 所示。方法共包含 3 部分，分别为聚类采样、ResNeXt-GRU 模块、Softmax 分类模块。首先，将原始行为视频经过提帧操作处理为图像帧序列，再通过聚类采样方法提取关键帧序列作为网络的输入。然后，使用 ResNeXt-GRU 网络模型提取行为的时空特征。为避免出现过拟合的情况，在 GRU 层融入 Dropout 技术，以提高网络的泛化能力和准确率。最后经过全连接层和 Softmax 分类器获得视频序列中行为的分类结果。

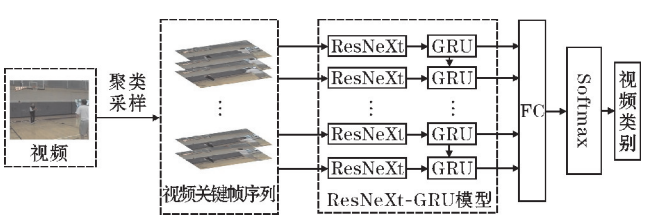

图 1 整体流程示意图

1.2 聚类采样

聚类算法[7]是机器学习中常见的一种算法。对于一个样本集，将其按照某种要求划分，然后以满足要求为目的不断迭代优化。它可以根据样本间的某种关系，优化样本质量，减少网络模型训练样本的规模，提

高最终的模型效果。

由于行为视频由连续的视频帧组成,因此,研究者在对人体行为视频特征提取前通常将视频处理成图像帧序列。但仅通过此操作处理后的视频帧序列,图像帧之间相似度很高,直接输入后续网络中进行训练,不仅增加训练时间,还影响最终的行为识别效果。因此,需要对图像帧序列做进一步的采样操作。LRCN 网络采用密集时间采样方式获取视频帧序列,通过将视频处理成图像帧序列,再随机获取其中一段等长连续的图像帧作为整段视频的表示。经过该采样方式获取的图像帧序列,帧间相似度高,不仅产生大量的冗余信息增加网络的计算成本,也存在容易丢失行为视频关键动作信息的风险。TSN 网络采用视频分段和稀疏采样方式对训练样本进行采样,先将完整视频进行分段,再从每段中选取一帧。相比 LRCN 网络,虽然一定程度上删减了冗余视频帧,但对于一些动作变换频繁的人体行为容易丢失关键帧。关键帧序列是视频中最具代表性的图像帧集合,该集合能够归结整段视频的中心内容。为充分提取视频中人体行为特征,本文基于聚类的方法对图像帧采样方式进行相应的改进。通过以图像帧间的相似度量为聚类标准,获取能够更好表示视频内容的视频关键帧序列作为网络的输入。实现步骤:(1)对视频进行帧采样为图像帧序列,并将图像帧的颜色空间由 RGB 转为 HSV,获取每帧图像的 HSV 直方图。(2)设置相似度阈值,以第一帧图像为初始聚类中心。(3)计算下一帧和每一个聚类中心的相似度,获取最大值。若小于阈值则自成一类;反之,加入此类,并重新计算聚类中心。重复此过程,直到取完所有帧。(4)计算每个类中图像帧与聚类中心的相似度,获取相似度最高的图像帧序号,按序输出对应的图像帧,并保存作为视频关键帧。本文关键帧提取的算法流程如图 2所示。行为视频在经过关键帧算法提取后所得到的为该视频的所有关键帧,但由于每个视频的长度和内容复杂情况不一致,最后得出的关键帧的数量也不同。本文固定长度为 k 的关键帧序列的取法为如果该视频的关键帧序列帧数少于 k,则取最后一帧补充;反之,在 0 到该关键帧序列总帧数 M 与 k 的差值间取一个整数,然后将这个整数作为片段的起始帧数并往后取连续的 k 帧作为选定好的片段。为验证聚类采样方法的性能,从数据集中任选一个视频,对其进行聚类采样操作,图 3 为一打篮球行为原始视频的全部图像帧,图 4 为该视频经过基于聚类的关键帧提取算法后所得的全部图像帧。通过对比图 3 和图 4 可知,视频帧序列在经过聚类采样操作后,序列中相似度高的数据样本得到了删减,一定程度上减少了数据规模,而保留下来的视频帧也能很好地表示视频中的人体行为。因此,通过对训练样本进行聚类采样处理,在保证样本集数据质量的同时,减少了数据规模,为后续网络的训练奠定了基础。

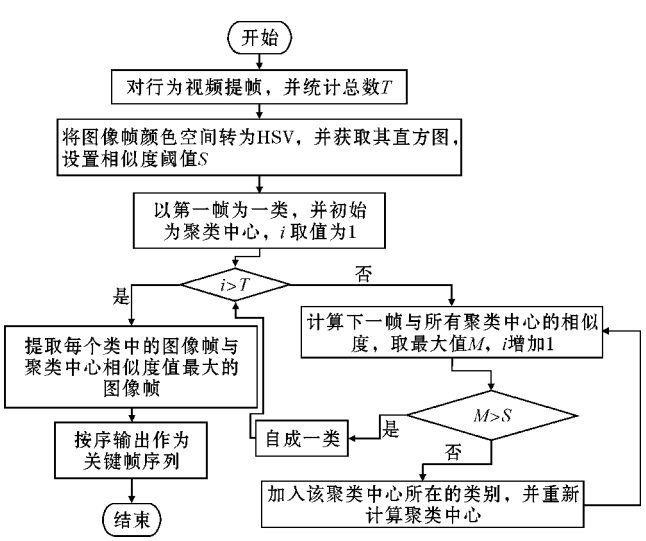

图 2　关键帧提取算法流程图

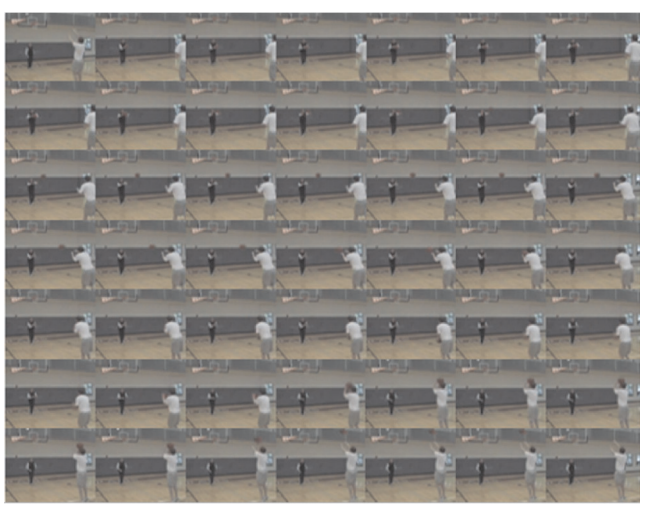

图 3　原始视频全部图像帧

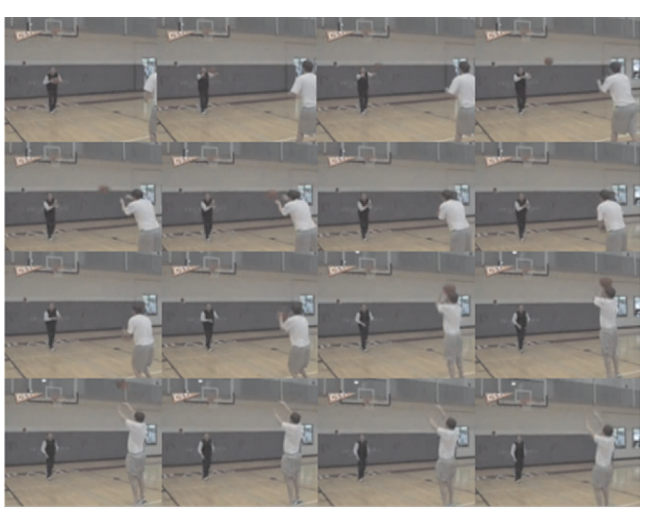

图 4　关键帧提取后视频全部图像帧

1.3　ResNeXt

人体行为识别的前提是能够提取行为中的有效特

征。由于卷积神经网[8](convolutional neural network, CNN)在图像识别任务中有很好的效果,研究者开始将卷积神经网络应用到行为识别任务中。利用 CNN 进行图像识别任务时,只需将图像直接输入到网络模型中,省略了传统算法中的人工特征提取过程,降低了模型处理复杂度。与全连接神经网络的不同之处在于,CNN 利用多层神经网络和图像局部性的优点减少了大量参数,提高了模型训练速度。常见的卷积神经网络模型有 GoogLeNet、AlexNet、VGGNet 等。因此,文中同样将经过聚类采样后的训练样本先输入到卷积网络中提取人体行为的空间特征。

由 Donhue 等提出的 LRCN 模型采用的是 AlexNet 网络来提取人体行为的空间特征,该网络主要由卷积层和池化层交替组成,网络结构简单,无法充分学习人体行为特征,且对复杂的人体行为识别效果不佳。为提高模型准确率,通常使用加深网络层数或拓宽网络宽度的方式。然而普通网络结构的叠加与拓宽,不仅容易导致网络退化,而且网络模型的参数也会大量增加。2017 年,Xie 等[9]提出了 ResNeXt 网络,它在 ResNet[10]网络的基础上集成了 VGGNet 网络堆叠和 Inception 网络拆分-转换-合并的思想,不仅能够解决网络退化问题,而且可以在不增加参数数量的前提下提高网络性能。因此,本文将 ResNeXt 网络作为提取人体行为空间特征的基础网络。ResNeXt 在 ResNet 网络的基础上优化而来。它的其中一个基本模块结构如图 5 所示,保留了 ResNet 中堆叠的 Block,不同之处在于 ResNeXt 将单个路径进行拆分,每个路径都为相同的拓扑结构,在每个拓扑结构都经过降维-变换-升维操作后再进行求和汇总。可用式(1)表示。

$$Y=X+\sum_{i=1}^{C}T_i(X) \tag{1}$$

式中,X 表示输入;Y 表示函数输出;T_i为相同的拓扑结构;C 为基数,表示一个模块中所具有的相同分支的数量,可以为任意数。实验表明,增加基数是获得精度的一种更有效的方法,文中 C 的取值为 32。

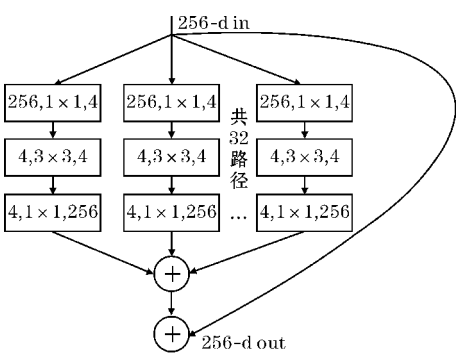

图 5 ResNeXt 基本结构图

1.4 门限循环单元(GRU)

人体行为视频具有时间属性,视频中的行为动作之间也往往具有一定的关联性。通过卷积神经网络虽然能够提取视频图像帧中的有效特征,但无法挖掘各图像帧间的时空上下文信息。为充分利用视频的时间维度信息,学习信息之间的依赖关系,本文在方法中加入 GRU[11]网络,提取视频中人体行为的时序特征。

GRU 是 LRCN 模型中 LSTM 网络的一种变体,功能与 LSTM 相同,但其结构更加清晰简洁,没有冗余结构。更少的参数也让其更不容易产生过拟合现象。GRU 网络模型单元结构如图 6 所示,主要包括一个更新门和一个重置门。其中,Rt 是 t 时刻的重置门,用于决定是否忘记之前的计算状态;Zt 是 t 时刻的更新门,用于控制将历史信息带入候选状态的程度。

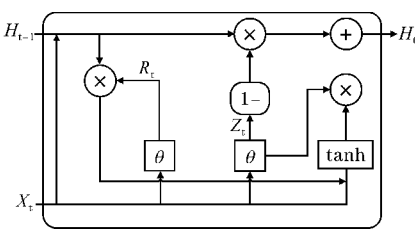

图 6 GRU 结构图

GRU 结构的计算公式如下:

$$\begin{cases} Z_t=\theta(W_zX_t+V_zH_{t-1}+b_z) \\ R_t=\theta(W_rX_t+V_rH_{t-1}+b_r) \\ C_t=\tanh(W_cX_t+V_c(R_t\odot H_{t-1})+b_c \\ H_t=(1-Z_t)\odot H_{t-1}+Z_t\odot C_t \end{cases} \tag{2}$$

式中,X_t 为当前 t 时刻的输入,C_t 为 t 时刻的候选状态,H_t 为 t 时刻隐藏层状态,H_{t-1}为 t 时刻之前的隐藏层状态,θ 为 sigmoid 逻辑函数,作为更新门和重置门的激活函数,tanh 为候选状态的激活函数,$\odot$为点积操作,W_z、W_r、W_c、V_z、V_r、V_c 为权重参数,b_z、b_r、b_c 为偏差参数。

由于直接使用全连接层进行特征融合,会导致高层特征无法捕获到空间特征在时域上的信息。因此本文使用 GRU 网络对 CNN 最后一层输出的卷积特征进行融合以获取 CNN 输出特征的上下文信息。

1.5 Softmax 分类器

行为识别本质上是一个多分类问题,文中当输入一个行为视频到网络后,需要判别视频中的内容属于 N 种行为中的哪一种。因此,在经过 ResNeXt-GRU 网络模型提取视频中人体行为时空特征后,再通过全连接层对所有输入值进行平均操作,最后使用 Softmax 分类器对所提取特征数据进行处理,进而完成人体行为的识别。Softmax 函数的定义式为

$$\text{Softmax}\ (S_j)=\frac{e^{S_j}}{\sum_{k=1}^{N}e^{S_k}} \tag{3}$$

式中,S_j 为分类器前的全连接层的输出,j 表示类别序

号,N 为总类别个数。

2　实验

2.1　数据集

2.1.1　数据集分析

本文在 UCF101 和 HMDB51 两个主流人体行为视频数据集上进行实验。UCF101 数据集是从 YouTube 网站上收集而来的,共包含 13320 个人体行为视频片段,每个视频片段持续 3 ~ 10 s,平均为 100 ~ 300 帧,分辨率为 320×240。它包括 101 个动作类,每类动作均由 25 人完成,每人做 4 ~ 7 组,其部分动作视频图像帧如图 7 所示。HDMB51 数据集共有 51 种类别,包含 6799 个视频片段。每个动作至少包含 51 个视频,分辨率为 320×240,来自于 YouTube,Google 视频等,包含单人行为、面部表情和操纵对象行为、人与人交互的行为、人与物交互等类别。部分动作视频图像帧如图 8 所示。

图 7　UCF101 数据集图像帧展示

图 8　HMDB51 数据集图像帧展示

2.1.2　数据集处理

实验选取每个数据集的 80% 作为训练集,20% 作为测试集,并将训练集和测试集按官方提供的方式划分成 3 组。针对数据集中的行为视频,首先将视频处理成图像帧序列,然后通过上文所提的聚类采样方法获取 k 帧图像序列,最后对长度为 k 的帧序列做以下同样的数据处理操作。具体操作:(1)为满足网络的输入大小,将图像帧分辨率由 320×240 处理为 224×224。(2)对训练集内的帧序列在空间上做上下左右的随机翻转,扩充数据的多样性。(3)为加快网络的训练和收敛,对帧序列进行归一化操作。

2.1.3　评价指标

为能对所提方法的表现进行评估和比较,本文使用基础的分类 Top-1 准确率作为评判标准来评估方法的准确程度,如式(4)所示。

$$\text{accuracy}=\frac{n}{N} \tag{4}$$

式中,n 为分类准确的样本数,N 为总样本数。

2.2　实验设置

实验环境基于 Window10 系统,Interl(R) Xeon(R) CPU,内存64 G,显卡为 NVIDIA TITAN Xp。实验所采用的深度学习框架为 Pytorch,集成开发环境是 Pycharm。

采用 ImageNet 上预训练的 ResNeXt101 模型对参数进行初始化,并使用随机梯度下降法(SGD)对网络进行训练,网络输入的批量数据(batch-size)大小设置为 16,动量设置为0.9,GRU 中的 dropout 设为0.5,学习率设置为0.0015,共迭代 120 个周期,损失函数使用交叉熵损失函数。

2.3　实验结果与分析

2.3.1　网络关键帧输入数量 k 对识别结果的影响

网络输入数据的大小对网络性能有重要影响,本文在将输入数据进行聚类采样后得到是每个视频的关键帧序列,由于视频内容的复杂情况不同,导致提取到的关键帧序列的长度也不一致。因此获取 k 的合适取值对于实现良好的识别效果至关重要。本文根据数据集聚类采样结果,将网络输入的关键帧序列长度 k 分别为 5,10,15,20,25,30,并输入到 ResNext101-GRU 网络,在其他实验条件一致前提下进行实验,并统计识别的准确率,实验结果如图 9 所示随着 k 值的增加,行为识别的准确率也随之增加,当 k 为 15 时,识别效果最好。之后,随着 k 值的继续增加,准确率却提高不大。考虑到输入网络的数据规模越大,不仅会增加网络的计算负担,而且效率变低。因此,本文最终确定 k 的为 15。

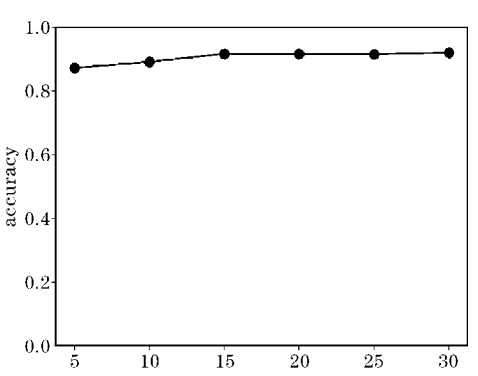

图 9　不同 k 值下识别结果对比

2.3.2 数据采样方式对识别结果的影响

不同的数据采样方式获得的训练样本存在一定的差异,影响着最终的识别结果。本文分别使用的 3 种数据采样方式:密集时间采样、视频分段采样、聚类采样,对数据集 UCF101 中 13000 多个视频均采样 15 帧,然后输入到 ResNeXt101-GRU 模型上进行实验。实验结果见表 1 可知,3 种采样方式中,训练样本采样时间用时最少的是密集时间采样方式,但识别效果却最低。采样时间用时最多的是聚类采样,平均每个视频所消耗的时间比视频分段采样方式只多了0.042 s,但识别效果最佳,平均准确率比密集时间采样高 4.4%,比视频分段采样方式高2.2%。实验证明了聚类采样方法的有效性。

表 1 不同数据采样方式对识别效果的影响

方法	采样耗/s	UCF101/%
密集时间采样	2437	89.3
视频分段采样	3127	91.5
聚类采样	3678	93.7

2.3.3 网络模型深度对识别结果的影响

除了网络输入数据的规模和采样方式,网络模型的深度也同样对识别结果有影响。在 k 为 15 的前提下,本文使用不同深度的空间特征提取网络 ResNeXt 在数据集 UCF101 上进行实验。实验结果见表 2可知,在 15 帧视频关键帧输入的 ResNeXt-GRU 模型中,随着模型网络层数的增加,网络模型的表征能力加强,行为识别的准确率越来越高。然而增加网络的层数,也会加大网络模型的运算量和运行时间。因此,综合考虑,本文确定 ResNeXt101 作为行为视频空间特征提取的网络模型。

表 2 不同网络模型深度对分类准确率的影响

方法	UCF101/%
ResNeXt18-GRU	84.3
ResNeXt34-GRU	88.6
ResNeXt50-GRU	92.2
ResNeXt101-GRU	93.7

2.3.4 本文方法与现有的主流方法的性能对比

基于以上实验结果,通过聚类采样操作获取每个行为视频的 15 帧关键帧序列,并输入组合模型 ResNeXt101-GRU 中进行实验,图 10 为该网络在 UCF101 和 HMDB 两种数据集上训练时 loss 值的下降曲线。保存实验结果,与现有主流方法在 UCF101 和 HMDB51 数据集上的平均识别率进行比较。实验结果如表 3 所示,对于 UCF-101 数据集,本文所提出的模型相对于目前识别效果最好的传统方法 IDT[12] 而言,准确率提高了 7.8%;与基于双流网络的方法 Two-stream CNN 和 TSN 网络相比,准确率分别提高了 5.7% 和 0.2%;相比基于三维卷积神经网络的经典方法如 C3D、P3D[13]、Res3D[14],本文方法行为识别的准确率更高;与基于 LSTM 的 LRCN 算法相比,准确率提高了9.8%;与蒋圣南等[15]提出的方法相比,虽然同样使用了 ResNeXt 网络,同样仅输入 RGB 图像这一种模态数据下,本文方法的准确率提高了5.9%。与文献[16]相比,模型结构与本文相似,都是结合 CNN 和 RNN 来识别人体行为,不同之处在于其使用三维卷积神经网络提取行为空间特征,实验结果表明,本文的准确率比其高0.08%。对于 HMDB51 数据集,识别效果虽然不如 UCF101 数据集,但同样优于大部分方法。

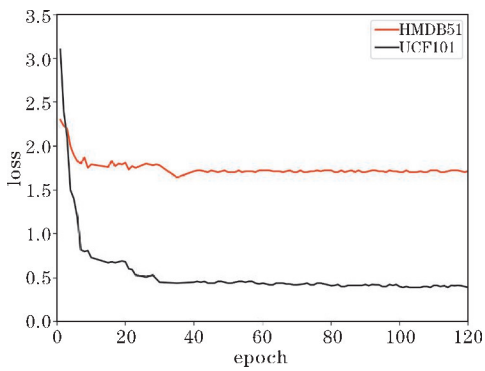

图 10 训练 loss 曲线图

表 3 不同方法在 UCF101 和 HMDB51 数据集上的识别准确率

单位:%

文献	方法	UCF-101	HMDB-51
文献[12]	IDT	85.9	51.9
文献[2]	Two-stream CNN	88.0	59.4
文献[3]	Two-stream CNN	93.5	68.5
文献[4]	3DCNN	82.3	–
文献[13]	3DCNN	88.6	61.2
文献[14]	3DCNN	85.8	54.9
文献[5]	2DCNN+LSTM	82.9	–
文献[15]	ResNeXt(RGB)	87.8	–
文献[16]	3DCNN+LSTM	93.62	–
本文	ResNeXt101+GRU	93.7	69.2

3 结束语

在 LRCN 模型的基础上,提出一种基于 ResNeXt-GRU 的人体行为识别方法。利用聚类算法改进网络输入数据的采样方式,减少冗余数据输入,提高识别效果。同时使用 ResNeXt 网络结合具有记忆功能的 GRU 网络,加强对视频中人体行为时空特征的提取。通过各种实验确定该方法最佳的输入视频帧数、采样方式和网络模型深度,在 UCF101 和 HMDB51 数据集上分别取得了93.7% 和69.2% 的准确率,与现有许多行为识别网络相比,准确率更高,说明了本文方法的有效性和可比较性。

参考文献:

[1] 凌佩佩,邱崧,蔡茗名,等. 结合特权信息的人体动作识别[J]. 中国图象图形学报,2017,22(4):482-491.

[2] Simonyan K,Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. arXiv preprint arXiv:2014,1406:2199.

[3] Wang L,Xiong Y,Wang Z,et al. Temporal segment networks:Towards good practices for deep action recognition[C]. European conference on computer vision. Springer,Cham,2016:20-36.

[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. 2015:4489-4497.

[5] Donahue J,Anne Hendricks L,Guadarrama S,et al. Long-term recurrent convolutional networks for visual recognition and description[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:2625-2634.

[6] Zhao H,Jin X. Human action recognition based on improved fusion attention cnn and rnn[C]. 2020 5th International Conference on Computational Intelligence and Applications(ICCIA). IEEE,2020:108-112.

[7] WU X iru,XUE Ganggang. 基于图像聚类的交通标志 CNN 快速识别算法[J]. 智能系统学报,2019,14(4):670-678.

[8] Hu H,Yang Y. A Combined GLQP and DBN-DRF for Face Recognition in Unconstrained Environments[C]. 2017 2nd International Conference on Control,Automation and Artificial Intelligence(CAAI 2017). Atlantis Press,2017:553-557.

[9] Xie S,Girshick R,Dollár P,et al. Aggregated residual transformations for deep neural networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:1492-1500.

[10] He K,Zhang X,Ren S,et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:770-778.

[11] Cho K,Van Merriënboer B,Gulcehre C,et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:2014,1406:1078.

[12] Wang H,Schmid C. Action recognition with improved trajectories[C]. Proceedings of the IEEE international conference on computer vision. 2013:3551-3558.

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

[14] Tran D,Ray J,Shou Z,et al. Convnet architecture search for spatiotemporal feature learning[J]. arXiv preprint arXiv:2017,5038:1708.

[15] 蒋圣南,陈恩庆,郑铭耀,等. 基于 ResNeXt 的人体动作识别[J]. 图学学报,2020,041(002):277-282.

[16] 陈颖,来兴雪,周志全,等. 基于 3D 双流卷积神经网络和 GRU 网络的人体行为识别[J]. 计算机应用与软件,2020,37(5):164-168,218.

Human Behavior Recognition based on ResNeXt-GRU and Cluster Sampling

ZENG Qingxi,　PENG Hui

(College of Software Engineering,Chengdu University of Information Technology,Chengdu 610225,China)

Abstract:In order to effectively capture the temporal relationships in behaviors and enhance the feature representation ability of the network, a human behavior recognition method based on ResNeXt-GRU is proposed. First of all, the behavioral video key frame sequences are extracted by using a clustering algorithm and then input to the ResNeXt network for feature extraction in spatial dimension. Then, the output feature vectors are all input into the gate recurrent unit (GRU) network for temporal learning. Finally, a Softmax classifier is used for classification. Experiments on UCF101 and HMDB51 datasets recognition accuracy of 93.7% and 69.2%, respectively. The experimental results show that the recognition accuracy was improved compared with many other existing behavior recognition methods.

Keywords:action recognition; clustering; ResNeXt; gate recurrent unit (GRU); Softmax