基于视频关键帧提取的快速 T3D 动作识别模型

丁建立　袁梓瑞*　王怀超
（中国民航大学计算机科学与技术学院　天津 300300）
（中国民航信息科研基地　天津 300300）

摘　要　视频级动作识别存在着数据量大、识别速度慢的问题，主要原因是需要提取空间维度上人体姿态，还需要考虑时间维度上动作关联。提出一种基于视频关键帧提取的快速 T3D 动作识别模型，通过改进的 Superpoint 网络提取视频关键帧，缩减数据量。以 T3D 网络为基础，时空分解其关键模块可变时序卷积层，显著提升了其计算效率。在公共数据集 UCF-101 和 HMDB-51 数据集进行了实验验证，准确率和原 T3D 网络近似，但其识别速度为原 T3D 网络的 2 倍，更适合于实际的应用场景。

关键词　快速动作识别　视频关键帧提取　T3D 网络　Superpoint 网络　快速识别

中图分类号　TP3　　**文献标志码**　A　　DOI:10.3969/j.issn.1000-386x.2024.08.026

FAST T3D ACTION RECOGNITION METHOD BASED ON VIDEO KEY FRAME EXTRACTION

Ding Jianli　Yuan Zirui*　Wang Huaichao
(*College of Computer Science and Technology*, *Civil Aviation University of China*, *Tianjin 300300*, *China*)
(*The Research Base of Civil Aviation Information Scientific of China*, *Tianjin 300300*, *China*)

Abstract　The video level action recognition method has the problems of large amount of video input data and slow recognition speed. The main reason is that these methods not only need to extract human posture in the spatial dimension, but also need to consider the association of actions in the temporal dimension. This paperproposes a fast T3D action recognition method based on video key frame extraction. It extracted video key frames through improved Superpoint network to reduce the amount of video data. Based on T3D network, the computational efficiency was improvedthrough spatiotemporal decomposition of its key module variable timing convolution layer. Experimental validation was conducted on the public datasets UCF-101 and HMDB-51. This method's accuracy is similar to the original T3D network, but its recognition speed is twice that of the original T3D network, which is more suitable for practical application scenarios.

Keywords　Fast action recognition　Key frame extraction　T3D network　Superpoint network　Fast recognition

0　引　言

动作识别是计算机视觉领域一个重要分支，由于其在安全保卫，视频分析等领域应用前景较为广阔，国内外学者对其开展了大量的研究，结合机器学习与深度学习模型的优良性能，视频级的动作识别准确率逐年攀升。但是动作识别模型普遍存在识别速度慢的问题。主要原因是动作识别模型需要提取视频帧动作随时间变化特征，而且一部分模型需要额外提取视频的光流图，导致最终输入模型数据量远大于原视频。动作识别模型时间特征提取模块参数多，模型优化困难，计算时间长，不符合快速识别的实际需求。

卷积神经网络由于其易于搭建，提取图像特征速度快的特点，广泛用于动作识别任务。主流的动作识别模型可以分为三类。第一类是双流网络，采用两个卷积网络来对视频的时间以及空间特征分别进行提取。Simonyan 等[1]提出了初始的双流网络用来进行

收稿日期：2021－03－01。国家自然科学基金项目(U1833114)；民航安全能力项目(SA2020280)；中央高校基本科研业务费项目(3122019120)。丁建立，教授，主研领域：民航智能信息处理与航空互联网。袁梓瑞，硕士生。王怀超，讲师。

动作识别,空间网络输入单帧图片,时间网络输入光流图序列,两个网络的输出经过融合后采用多分类的 SVM 分类器输出分类结果。Munro 等[2]训练了一个自监督分类器用来强化双流网络对不同场景的同一动作的识别效果,提高双流网络的泛化性能。Li 等[3]提出两种不同时间跨度的模块,短时模块利用 2D 卷积提取帧特征并进行时间维度的池化,长时间模块先对图像帧序列进行分组,对每一组进行卷积提取以组为时间单位的行为特征。网络提升了动作识别的准确率,但是识别速度较慢。第二类是 3D 网络,初始的 3D 网络模型是由 Ji 等[4]提出的,通过三维卷积层从空间和时间两个维度提取特征,捕获特征在多个相邻帧中的运动信息进行视频级的动作识别。Diba 等[5]也仿照 2D 卷积网络中的 DenseNet 提出了 T3D 模型,通过构建多尺度时间层模拟可变时间卷积,改善了 3D 网络不能利用长时间动作特征的缺点。Tran 等[6-7]借鉴 R3D 模型,尝试用 2D 卷积层代替了某些 3D 卷积层,以简化对时间不敏感的层,并将时间维度添加到了 2D 卷积层的通道深度里。同时拆分 3D 卷积核,以减少参数量。Zhou 等[8]提出了 2D/3D 跨域残差并联模块,克服了 3D 网络容易梯度消失,网络模型不能过深的问题。Feichtenhofer 等[9]受 3D 卷积对于时间维度扩展提升动作识别精度的启发,以 R3D 模型为基础,运用特征选择方法,分别对视频帧速率、长度、分辨率、网络通道数和层数进行扩展并进行对比。最终大幅降低了网络参数,但是对于待识别视频要求较高,且训练难度较大。第三类是长期递归神经网络,简称 LRCN,运用了长短期记忆网络(LSTM)对可变长度的视频帧序列进行处理,相比于 3D 网络更容易反向传播优化。Donahue 等[10]提出了原始的 LRCN 模型,该模型用 CNN 提取出图像的空间特征,与可以学习识别和合成涉及顺序数据的任务或时间动态模型 LSTM 结合,提升了长时间跨度动作的识别精度。Wang 等[11]提出了一种新的时空序列学习的网络结构 E3D-LSTM。通过引入了一种回溯机制来改善当前单元记忆状态,即当前单元状态取决于前一步记忆和过去某段时间的记忆的融合,该机制有助于平衡短期与长期时间特征对识别结果的影响。谢昭等[12]运用了滑动窗口方式自适应不同长度的视频,通过时空关注度来提升某些帧的关键区域对于识别结果的影响。

三类模型的缺点比较明显。双流网络需要提取额外的视频光流图,输入数据量过大导致普遍识别速度慢。3D 网络中的 3D 卷积层难以优化、参数量大、训练及识别速度偏慢。LRCN 则很难捕捉短时间的动作,其次是难以训练,容易导致模型参数不收敛。这三类模型的共同问题就是识别的速度偏慢,不符合实际需求中快速识别的要求。

对于视频数据,同一场景连续的视频帧中有很多冗余的信息。对于动作识别任务来说,每种动作都有关键步骤,对应着视频中动作的关键帧。通过提取关键帧可以有效地削减动作识别模型的训练时间及识别时间。Kar 等[13]将经过 CNN 提取的空间特征经过自适应池化层,结合到目前为止的视频向量推断当前帧特征的重要性,将关键帧赋予较大权重作为动作的关键帧。Mahasseni 等[14]利用 CNN 来生成原始视频的深度特征,GAN 网络不断提取关键帧重构原视频的深度特征,求得两个深度特征最接近的关键帧生成器。本文改进的是 Detone 等[15]的工作,该模型原本用于图像的无监督标注,经过改进用 CNN 编码网络分别提取图像特征点及描述子。由连续帧中描述子确定的对应特征点的欧氏距离来筛选出关键帧。

近年来针对动作识别网络的改进偏向于提高识别准确度,而识别速度往往不是研究的重点。针对动作识别时间长的问题,本文提出了一种快速动作识别模型,利用改进的 Superpoint 网络提取动作关键帧,缩减视频数据量规模。改进现有的 T3D 模型,时空分解其关键模块可变时序卷积层,进一步加快了识别速度。改善了现有动作识别模型识别速度慢的问题。

1　模型框架

1.1　整体架构

从数据量输入的角度来看,3D 卷积网络是最好的选择,只需要输入原始视频切分成的帧序列即可,但是数据量依旧很大,3D 网络要花费大量的时间和计算资源去拟合实际场景中的不同动作。且 3D 网络中的卷积层参数过多,优化困难导致训练和识别时间过长,模型收敛较慢。这些问题使得 3D 网络达不到实际场景中的识别速度需求。为解决上述问题,提出了一种快速动作识别模型,其架构如图 1 所示。

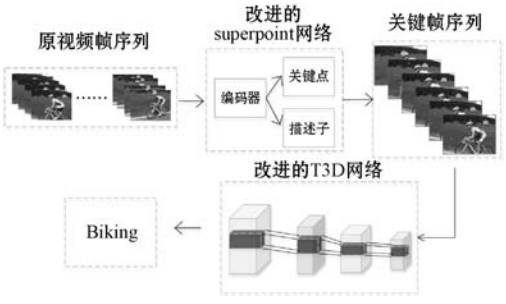

图 1　模型整体架构

由于视频数据量大,所以采用改进的 Superpoint 网络对原始视频进行动作关键帧的提取,通过编码器提取出每一帧的关键点与描述子。在不同帧之间,通过描述子确定的对应关键点的平均距离确定是否为关键帧。抽取出的关键帧数量为原视频帧的三分之一左右,显著降低了输入视频数据量的大小。同时改进 T3D 网络,将其关键模块可变时序的卷积层 TTL 做时空分解,简化了纯 3D 卷积层的运算,反向传播也分别进行,对提升网络训练速度及识别速度有帮助。

1.2 改进的 Superpoint 网络

如图 2 所示,网络由两部分组成:第一部分是特征点提取网络,用来提取出图像的角点(一种通用的特征点,包括三维物体的边和角);第二部分是另一个并行的描述子提取网络,提取出特征描述子,描述子用来唯一确定对应的特征点。提取关键帧的思路是将两帧图片用 Superpoint 分别提取特征点和描述子,由描述子匹配两帧图片中对应的特征点,计算两帧图像对应特征点的平均距离,再和事先确定的距离阈值作比较来确定是否为关键帧。

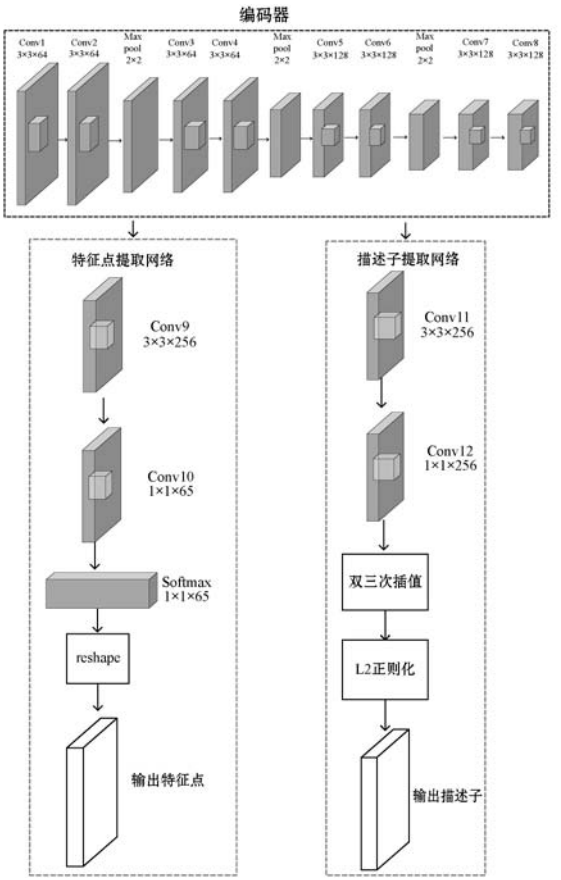

图 2 改进的 Superpoint 网络

编码器采用 VGG 形式的网络构成,由 8 层卷积层,3 层最大池化层构成,每个卷积层后均使用 ReLU 函数激活,每个卷积层卷积核大小都为,滑动步长为 1,池化核大小都为,滑动步长为 2,这种结构可以使特征图维度迅速下降,同时避免池化程度过深信息损失增大。随着网络的深入,卷积形成的特征图会越来越抽象,所以卷积核数量每隔四层卷积会翻倍。编码器输入图像 $\boldsymbol{I}\in\mathbf{R}^{H\times W\times F}$,$H$ 和 W 代表图像的宽和高,F 表示输入的维度,这里是 1。输出特征图为 $\boldsymbol{I}\in\mathbf{R}^{H_{\mathrm{d}}\times W_{\mathrm{d}}\times 65}$($H_{\mathrm{d}}=H/8$,$W_{\mathrm{d}}=W/8$)。

描述子提取网络与特征点提取网络共享同一个编码器,先由编码器将原图降维至 $\boldsymbol{I}\in\mathbf{R}^{H_{\mathrm{d}}\times W_{\mathrm{d}}\times 65}$,然后输出描述子的半密集网格(在原图的宽度上每 H_{d} 个像素,高度上每 W_{d} 个像素作为一个单位输出描述符,用来对应特征点提取网络输出张量中每一个特征点),描述子提取网络做双三次插值,L2 正则化,上采样,输出的描述子维度为 $1\times1\times256$,对应每个特征点在原图的网格位置,以及网格内的位置,由描述子可以找到对应的特征点。

关键帧提取步骤如图 3 所示。

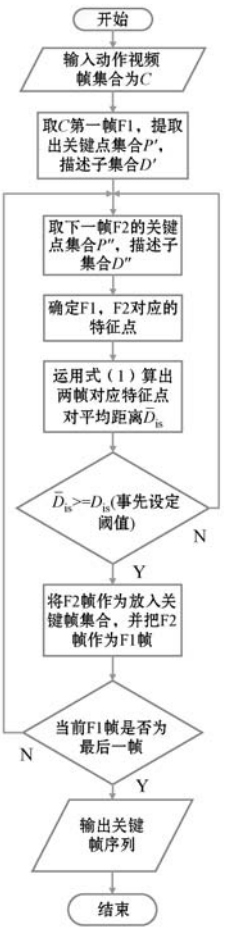

图 3 改进的 Superpoint 提取关键帧流程

利用描述子 $D'_i\in D'$ 所在的局部区域可以确定唯一的特征点 P'_i,同理在第二帧的同一区域利用描述子 $D''_i\in D''$ 可以确定唯一的特征点 P''_i,这样求得的 P'_i 和 P''_i 称为对应特征点对,如式(1)所示。

$$\overline{D}_{\mathrm{is}}=\sum_{i=1}^{64}\sqrt[2]{\left[P'_i(x,y)-P''_i(x,y)\right]^2}/64 \tag{1}$$

求得的 $\overline{D}_{is}$ 为对应特征点对的平均距离。

使用改进的 Superpoint 网络提取的关键帧结合了视频中全部画面的局部特征,能够对人体动作做出较大改变时,捕捉到身体各个部分的特征改变情况,提取出有助于识别动作信息的关键帧。在人体动作没有较大改变的连续帧序列,改进的 Superpoint 网络能够减少对这些帧的关注,避免提取出的关键帧过于密集,造成动作信息的冗余。

1.3　改进的 T3D 网络

T3D 网络是行为识别领域的一个经典模型,其最大特色就是创建了可变时序的卷积层 TTL,可以对 3D 特征图分别进行不同时间长度的卷积,提取出不同时间维度的特征。本文将其关键模块可变时序的卷积层 TTL 做时空分解,得到如图 4 所示的网络模型。

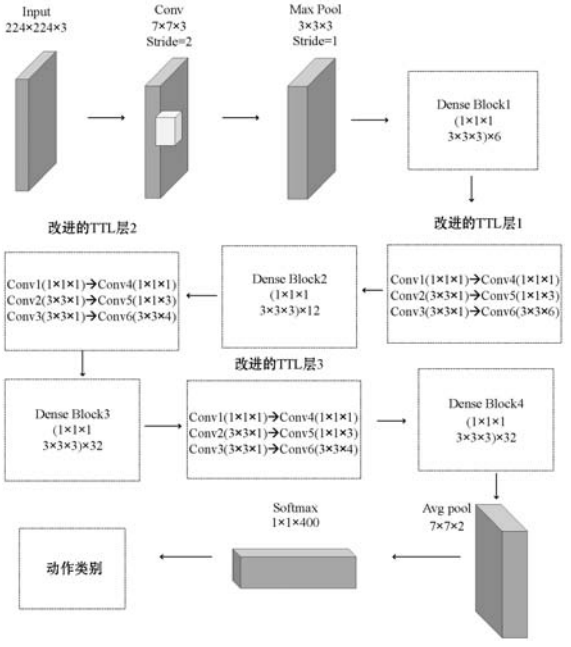

图 4　改进的 T3D 网络

网络输入图像尺寸大小为 224 × 224,每一个 Denseblock 采用密集连接。以 Denseblock1 为例(见图 5),每一个卷积组,即 1 ×1 ×1 的卷积层和 3 ×3 ×3 的卷积层,与其他的卷积组之间均有一条连接,6 个卷积组之间共有 15 条连接。

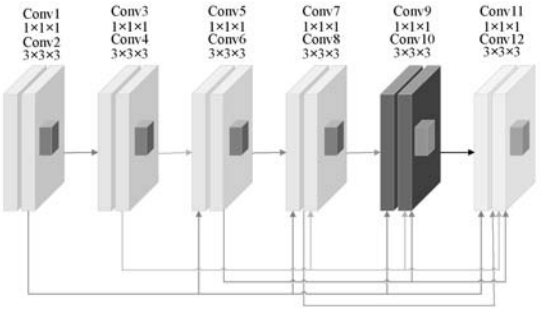

图 5　Denseblock1 结构图

网络共有四个 Denseblock 块,第一个块有 6 个卷积组,第二个块有 12 个卷积组,第三个块和第四个块都有 32 个卷积组。

原始的 T3D 网络中的 TTL 层利用三种不同时间维度的 3D 卷积核,来处理短、中、长三种时序信息的 3D 特征图,如图 6 所示。

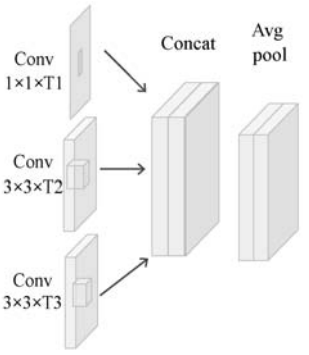

图 6　TTL 层结构示意图

TTL 层可用式(2)表示。

$$F = f(t_1 c_1, t_2 c_2, t_3 c_3) \tag{2}$$

式中:t_1、t_2、t_3 代表不同长度的时间维度卷积核,c_1、c_2、c_3 代表不同大小的空间维度卷积核。对于 3D Denseblock 输入的张量 $\boldsymbol{x}$,TTL 层的工作可表示为:

$$F(\boldsymbol{x}) = f(t_1 c_1(\boldsymbol{x}), t_2 c_2(\boldsymbol{x}), t_3 c_3(\boldsymbol{x})) \tag{3}$$

$$\boldsymbol{X}' = avgpool(concat(F(\boldsymbol{x}))) \tag{4}$$

式中:*avgpool* 代表平均池化;*concat* 代表张量拼接,经过三种不同时空维度卷积的输入张量 $\boldsymbol{x}$ 最终统一为张量 $\boldsymbol{X}'$ 输出。

但是,对于原始 3D 网络参数多,参数优化困难,训练以及识别时间长等问题并没有做出较大改善,T3D 的识别速度还达不到快速识别的要求。所以为了减少训练及识别时间,对 TTL 层中的 3D 卷积块做分解,如图 7 所示。

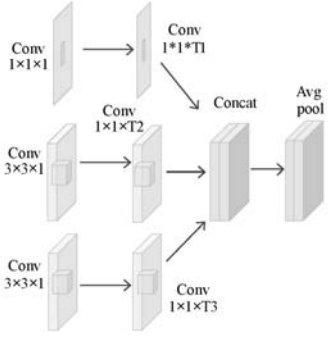

图 7　改进的 TTL 层

同样对于输入 $\boldsymbol{x}$,改进的 TTL 层可表示为:

$$F(\boldsymbol{x}) = f(t_1(c_1(\boldsymbol{x})), t_2(c_2(\boldsymbol{x})), t_3(c_3(\boldsymbol{x}))) \tag{5}$$

式中:$t(c(\boldsymbol{x}))$ 表示先经过空间卷积层,再经过时间卷积层。

把一个 3D 卷积核拆解成两个 3D 卷积核,分别提取空间特征和提取时间特征。对于空间卷积核,时间维度全部置为 1,等同于只做了 2D 空间上的卷积,而对于时间卷积核,空间维度均设为 1 ×1,等同于只做了 1D 时间上的卷积。这样做的好处减少了运算次数,以 $n \times n \times n$ 的卷积核为例,一次卷积所需要的计算次数为 $n \times n \times n = n^3$ 次,但是拆解成两个卷积核运算次数降为 $n \times n \times 1 + 1 \times 1 \times n = n^2 + n$ 次,降低了一个幂指数级。在 T3D 网络中,3D 卷积核的时间维度通道时通常大于空间维度尺寸大小,所以实际运算次数减少得更多。

改进的 TTL 层共有三个,第一个 TTL 层的中长时间 3D 卷积块空间维度大小均为 3 ×3,故空间维度共享 3 ×3 ×1 的卷积块,时间维度根据大小分解为 1 ×1 ×3 和 1 ×1 ×6 两个卷积块。第二个 TTL 层中长时间的卷积块空间维度也是 3 ×3 ×1,时间维度根据大小分解为 1 ×1 ×3 和 1 ×1 ×4 两个卷积块。第三个 TTL 层构造和第二个相同,所有 TTL 层的时空分解卷积块后都跟着一层张量拼接层。因为不同时间维度的卷积块输出的张量维度不同,最后均由 2 ×2 ×2 的平均池化层提取更深一层的特征。最后输出的张量大小为 7 ×7 ×2,通过平均池化和后由 Softmax 提取分类信息。

2 模型效率计算参数

2.1 关键帧提取参数

在网络实际处理连续帧时,将图像宽高压缩为实际的 1/3,默认的关键帧最大间隔为 10 帧,避免长时间静止场景信息被过滤掉。关键点距离阈值 Dis 设置为 0.7 个像素距离,可以使筛选出的关键帧在原视频中的分布较为均匀,关键帧的采样率设置为 35% ,提取出的关键帧数量约为原视频帧的 1/3,从数据源头减少了视频数据量。由于原始网络对于三维物体的轮廓提取效果明显,因此网络使用了原始 Superpoint 权重,当人体做出幅度较小的动作时,提取的关键帧在原视频中的分布依然比较均匀。有利于改进的 T3D 网络的训练及识别。

2.2 改进的 T3D 网络参数

改进后的 T3D 网络参数数量与原网络基本相同,不同之处在于经过改进的 TTL 层时空分解了 3D 卷积核,参数数量进一步降低,具体参数数量对比如表 1 所示。

表 1 改进前后三层 TTL 层参数数量

网络层	输出张量	原网络参数数量	改进后网络参数数量
TTL 层 1	28 ×28 ×8	1.34×10^6	0.46×10^6
TTL 层 2	14 ×14 ×4	$3.121\ 4 \times 10^5$	1.155×10^5
TTL 层 3	7 ×7 ×2	6.654×10^4	2.864×10^4

改进后的 3 个 TTL 层的参数数量约为原网络的 1/3,降低了运算量。改进后的 TTL 层每一个 3D 卷积层时空分解成两个子卷积层。两个子卷积层所提取的特征经过后面的非线性激活函数后拟合效果更好。最重要的一点是时空分解让反向传播优化的过程也分别进行,相比于纯 3D 卷积核,优化效果更加明显。

2.3 准确率参数计算

此模型评估指标为识别准确率,公式如式(6) -式(7)所示。

$$y_i = \begin{cases} 1 & \hat{y} = y_{\text{pred}} \\ 0 & \hat{y} \neq y_{\text{pred}} \end{cases} \tag{6}$$

$$A_{\text{accuary}} = \sum y_i / n_{\text{um}} \tag{7}$$

式中:$\hat{y}$ 是输出的 Softmax 向量中最大值的下标,y_{pred} 是真实值独热码为 1 的下标,如二者相同,则输出正确,此输出记为 1,否则记为 0。式(7)中 n_{um} 为测试集总数,输出的 A_{accuary} 为准确率。

3 实验对比分析

3.1 数据集选取

本文采用两个公共数据集,第一个是 UCF-101 数据集,视频均来源于 Youtube,分为 101 类动作,每个动作类的剪辑包含 25 组,每组分为 4 到 7 个剪辑,每个剪辑均具有的分辨率与 25 帧每秒的帧速率,共有 13 320 个剪辑,时常共 1 600 分钟,格式为 AVI,是目前动作识别领域最常用的数据集之一。

第二个是 HMDB-51 数据集,有 51 类动作,共有 6 849 个视频,每个动作至少包含 51 个视频,分辨率。帧速率 25 帧每秒。视频来源于 YouTube、Google 等,时长共计 1 200 分钟,格式为 AVI。

3.2 实验环境

实验环境有两个。第一个为 CPU:Intel Core i5-9300H,2.40 GHz;运行内存:8 GB;GPU:NVIDIA GeForce GTX 1650,8 GB 显存;操作系统:Windows 10 家庭中文版 64 位;编程语言:Python 3.7.7;加速环境:CUDA 10.1,cuDNN 7.6.5;深度学习架构 PyTorch 1.7.0。此环境

主要运行关键帧提取网络。第二个为 CPU:Intel Xeon Bronze 3104,1.70 GHz × 6;运行内存:16 GB;GPU:NVIDIA Quadro P4000;操作系统:Ubuntu 16.04LTS 64 位;编程语言:Python 3.7.7;加速环境:CUDA 10.0,cuDNN 7.5;深度学习架构 TensorFlow 2.0。此环境主要运行的是改进的 T3D 网络。

3.3　不同动作识别网络性能对比

对比其他的视频动作识别网络,双流网络处理光流图序列的子网络拉长了整个网络的训练时间,而且在实际应用时,需对采集到的视频进行光流图的提取,提取出的光流图质量决定了识别的精确度。而对于 3D 网络来说,纯 3D 的卷积层和池化层会严重影响运行速度,而且收敛比较慢。P3D 由于需要多 GPU 并行训练,所以单个 GPU 训练时间较长。T3D 输入数据量小,网络结构简单,且改善了大多数动作识别模型的通病,对于时间跨度较长的动作识别效果不佳的问题。而改进的 Superpoint 网络会提取出视频中动作的关键帧,会减少视频中无用信息和冗余信息。因此在处理相同数据量的视频时,动作识别的速度也会加快。如表 2 所示。

表 2　不同模型识别性能对比

网络	输入数据	FPS	UCF-101/%	HMDB-51/%
C3D	视频帧序列	8.6	79.6	53.5
P3D	视频帧序列	9.3	86.0	57.6
T3D-121	视频帧序列	9.5	85.2	58.3
T3D-169	视频帧序列	9.4	87.8	59.4
Two Stream	视频帧序列 + 光流图序列	8.8	84.6	56.8
改进的 Superpoint + T3D-169	关键帧序列	14.6	86.4	58.2
本文	关键帧序列	16.8	87.6	59.0

可以看出,用不同网络训练视频数据集 UCF-101 和 HMDB-51 的实验对比,输入图片大小调整为 224 × 224。训练取一个 batch 为 10,学习率均设为 0.0001,训练轮数都设定为 200 轮,优化策略为 Adam。测试集为数据集中随机抽取 20% 的视频,并打乱顺序。此处的处理速度是将网络的运行时间与数据视频帧相除得到。

可以看出,在基础网络中,T3D 网络识别的基础速度达到了 9.4 帧每秒,是处理速度最快的,而在 UCF-101 数据集和 HMDB-51 数据集上取得的准确率是基础网络最高的。因此选取 T3D 网络作为本文的基础模型进行改进。通过识别改进的 Superpoint 网络提取的关键帧序列,网络的处理速度由 9.4 帧每秒提升至 16.8 帧每秒,识别速度接近原网络的两倍,且识别准确度与原始 T3D 网络近似相等,达到了提升识别速度的要求。

3.4　网络收敛速度对比

通过将 T3D 网络的关键模块可变时序的卷积层的时空分解,使得原来反向传播优化困难的 3D 卷积层变成了相对易于优化的两个卷积层,对于提升模型的收敛速度,减少训练时间有很大的帮助。

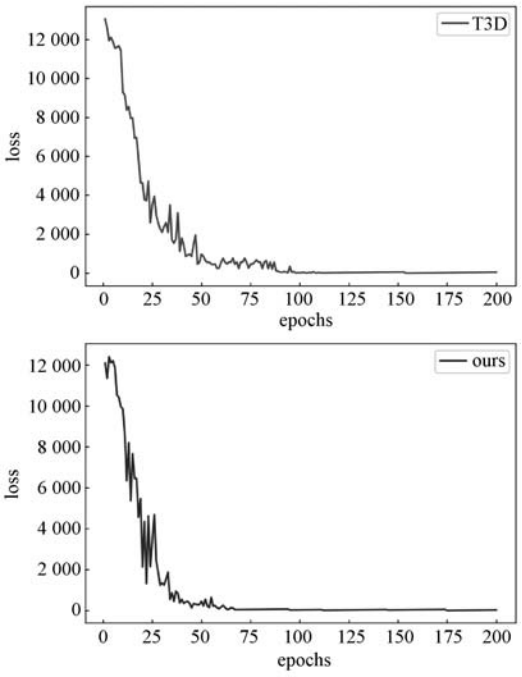

图 8　原 T3D 网络与改进之后的 T3D 网络收敛速度对比

此实验中,数据集为 UCF-101,学习率均设为 0.000 1,训练轮数都设定为 200 轮,优化策略为 Adam,选用交叉熵损失函数。可以看出,原 T3D 网络在接近 100 轮时 loss 值才趋于平稳,而改进之后的 T3D 网络在 70 轮左右 loss 值已趋于平稳,即改进的 T3D 网络收敛速度明显快于原网络。

3.5　不同类别动作的准确率对比

T3D 的良好特性之一是对长时间跨度的动作识别效果较好,TTL 层的不同时间维度卷积核可以良好适应不同时间跨度的动作,并不会出现动作分解的现象,而改进的 T3D 网络不仅在识别性能上显著提升,而且也继承了 T3D 网络适应长时间跨度动作的良好特性。

表 3 是对五类不同时间跨度动作,不同动作识别模型的准确率。这五类动作数据均选自 UCF-101。可以看出,T3D 网络对大部分动作识别准确率较高,尤其是对于跳高、板球等时间跨度大,动作复杂的动作识别

效果明显优于其他模型。改进之后的模型在损失精度较小的情况下,保留了原始 T3D 网络对长时间跨度的动作识别效果较好这一特性。

表 3 不同动作之间准确率对比(%)

动作类别	骑车	板球	爬行	跳高	弹吉他
C3D	90.6	86.8	78.6	70.7	83.5
P3D	92.5	88.3	80.5	75.7	87.3
Two stream	93.0	87.2	79.5	74.6	85.6
T3D	95.8	92.3	83.7	77.6	91.6
本文	95.2	92.0	83.5	77.3	91.4

4 结 语

针对现有的动作识别 3D 卷积网络模型输入数据量大,训练及识别速度慢的问题,本文从两方面出发,使用改进的 Superpoint 网络将视频数据压缩成关键帧序列,从数据源头减少网络训练及检测数据量。本文以 T3D 网络模型为基础,针对 3D 卷积层参数多、优化困难的缺点,改进了其关键模块可变时序的 3D 卷积层,使其降低了计算量。本文提出的提取关键帧模型结合改进的 T3D 网络,在公共数据集 UCF-101 和 HMDB-51 上取得了近似的精确度,而训练速度以及识别速度提升了 1 倍,且由于视频经过关键帧提取,减少了输入数据量,在实际应用中,更贴合快速识别的需求,同时可以降低视频存储空间。未来工作目标是将此模型运用到机场大厅旅客的动作实时识别,为此需要收集机场监控视频数据并扩充机场旅客特有动作类别,为机场的安保策略的制定提供帮助。

参 考 文 献

[1] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Advances in neural information processing systems, 2014, 27: 568 - 576.

[2] Munro J, Damen D. Multi-modal domain adaptation for fine-grained action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2020: 122 - 132.

[3] Li Y, Ji B, Shi X, et al. Tea: Temporal excitation and aggregation for action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE,2020: 909 - 918.

[4] Ji S, Xu W, Yang M, et al. 3D Convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1): 221 - 231.

[5] Diba A, Fayyaz M, Sharma V, et al. Temporal 3D convnets: New architecture and transfer learning for video classification[EB]. arXiv:1711.08200, 2017.

[6] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. IEEE,2018: 6450 - 6459.

[7] Tran D, Ray J, Shou Z, et al. Convnet architecture search for spatiotemporal feature learning[EB]. arXiv:1708.05038, 2017.

[8] Zhou Y, Sun X, Zha Z J, et al. Mict: Mixed 3D/2D convolutional tube for human action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2018: 449 - 458.

[9] Feichtenhofer C. X3D: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 203 - 213.

[10] 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. IEEE,2015:2625 - 2634.

[11] Wang Y, Jiang L, Yang M H, et al. Eidetic 3D LSTM: A model for video prediction and beyond[C]//International Conference on Learning Representations,2018.

[12] 谢昭,周义,吴克伟,等. 基于时空关注度 LSTM 的行为识别[J]. 计算机学报,2021,44(2):261 - 274.

[13] Kar A, Rai N, Sikka K, et al. Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 3376 - 3385

[14] Mahasseni B, Lam M, Todorovic S. Unsupervised video summarization with adversarial LSTM networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE,2017: 202 - 211.

[15] Detone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE,2018: 224 - 236.

[16] 郭洪涛,龙娟娟. 基于深度神经网络和投影树的高效率动作识别算法[J]. 计算机应用与软件,2020,37(4):273 - 279.