MRTP:时间-动作感知的多尺度时间序列实时行为识别方法

张坤[1]，杨静[1,2]，张栋[1]，陈跃海[1]，李杰[1]，杜少毅[2]
(1. 西安交通大学自动化科学与工程学院，710049，西安；2. 西安交通大学人工智能学院，710049，西安)

摘要：针对行为识别中时空信息分布不均衡以及对长时间跨度信息表征获取难的问题，提出了一种时间-动作感知的多尺度时间序列实时行为识别方法 MRTP。以 RGB 视频为输入，使用两个并行的感知路径在不同的时间分辨率上对视频进行空间特征与动作特征提取。在空间路径中，使用基于特征差分的动作感知寻找并加强通道动作特征表征；在动作路径中，基于动作感知的权重对通道进行筛选，并加入通道注意力和时间注意力加强关键特征；在两个路径提取出特征后，对特征进行融合，融合后的特征通过激活函数映射出样本在各个类别的得分，取得分最高的类别为最终识别结果。实验结果表明：所提方法在 UCF101 数据集上达到了 95.6%的准确率，优于未使用时间注意力的方法；在 AVA2.2 数据集上的平均精度达到了 28%，优于未使用动作感知和时间注意力的方法。与目前主流的基于光流法的双流网络、以 Slowfast 为代表的 3D 卷积网络、Transformer 等方法进行了准确率、参数量、处理速度对比，结果表明所提方法具有更良好的识别效果和鲁棒性。

关键词：行为识别；双路径网络；特征差分；动作感知；时间注意力
中图分类号：TP391.4 **文献标志码**：A
DOI：10.7652/xjtuxb202203003 **文章编号**：0253-987X(2022)03-0022-11

OSID 码

MRTP: Multi-Temporal Resolution Real-Time Action Recognition Approach by Time-Action Perception

ZHANG Kun[1], YANG Jing[1,2], ZHANG Dong[1], CHEN Yuehai[1], LI Jie[1], DU Shaoyi[2]
(1. School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
2. College of Artificial Intelligence, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: In view of the uneven distribution of temporal and spatial information and the difficulty in obtaining long-term information representation, we propose a dual-path action recognition method MRTP based on time and motion perception, which takes RGB video as input and applies two parallel perception paths to extract spatial and motion features from the video at different time resolutions. In the spatial path, the action-perception based on feature difference is applied to find and strengthen the action feature representation; in the action path, the channel is filtered based on the weight of action perception, and channel attention and time attention are added to enhance the key features; the characteristics of the two paths are merged to calculate the action category score of the video. The experimental results show that MRTP achieves accuracy of 95.6% on the UCF101 data set to outperform the model without time attention; on the AVA2.2 data set, the accuracy of mAP reaches 28% to outperform the model without action perception

收稿日期：2021-06-21。 作者简介：张坤(1997—)，男，硕士生；杜少毅(通信作者)，男，教授。 基金项目：国家自然科学基金资助项目(62073257)。
网络出版时间：2021-10-03 网络出版地址：http://kns.cnki.net/kcms/detail/61.1069.T.20210930.1331.002.html

and time attention. Compared with the current mainstream two-stream network, 3D convolution, Transformer, and other methods on a number of accuracy indicators, this method is endowed with better recognition effect and robustness.

Keywords: action recognition; dual-path network; feature difference; action perception; temporal attention

近年来，行为识别在智能视频监控、辅助医疗监护、智能人机交互、全息运动分析及虚拟现实等领域均具有广泛的应用需求[1]。从应用场景看，行为识别可分为异常行为识别、单人行为识别、多人行为识别等[2]。行为定义模糊、类内和类间差异较大、计算代价等问题给视频行为识别带来了巨大的挑战[3]。

随着深度学习的崛起，许多深度学习方法被用于行为识别。由于行为识别需要同时获取空间和时间信息，所以两个网络并行的双流结构成为了目前视频行为识别领域的主流架构。双流网络大多使用光流作为时间流、RGB 图像作为空间流。由于光流本身只使用于短时间的动作信息提取，所以此类网络无法解决长跨度动作的时间信息提取问题[4]。

循环神经网络在序列数据的处理上表现优异，而视频也是按照时序排列的序列数据，所以诸如 LSTM[5] 等循环神经网络被用于视频行为识别任务。然而，使用 CNN-LSTM 的方法在行为识别问题上并不能取得令人满意的效果。原因在于行为识别中作出主要贡献的是帧图像的空间信息[6]，且相邻的视频帧能提供的时序信息十分有限。

3D 卷积相较于 2D 卷积多了一个维度，对应视频比图像多了时间维度，因此 3D 卷积被引入用作行为识别的特征提取。随着视频领域大规模数据集的建立，3D 卷积逐步超越了传统 2D 卷积的表现[7]。然而，视频信息在时空维度具有完全不同的分布方式和信息量，经典的 3D 卷积方法在时空维度并没有对此进行区分[8]，由此导致了 3D 卷积计算了过多的冗余信息。如何减少 3D 卷积的计算消耗从而建立一个轻量级的网络是目前的研究热点。

长跨度的时间建模是行为识别中的一大难点[9]。由于时间维度信息与空间信息不平衡，已有的行为识别方法受限于采样密度较低和时间跨度限制，对于一些变化缓慢或者变化较小动作，如倾听、注视、打电话等，难以提取出有效的动作信息。对于部分需要依赖时间信息进行区分的动作，如讲话和唱歌、躺下和睡觉等，已有方法的效果不够理想。如何从冗余的视频信息中找到出含有动作信息的关键视频帧，目前的行为识别方法还未给出一个完善的解决方案。

本文针对 RGB 视频的轻量行为识别，提出了一种时间-动作感知的多尺度时间序列实时行为识别方法 MRTP，旨在解决视频中空间和时序信息不平衡以及长时动作的关键帧难以提取的问题。本文提出的 MRTP 方法在行为识别的经典数据集 UCF-101 和大规模数据集 AVA2.2 上进行了训练和相关指标测试。测试结果表明，相比于主流的行为识别方法，MRTP 方法具有更高的准确率和更小的计算成本，能够在方法部署阶段实现实时行为识别。

1　相关工作

行为识别传统方法一般使用时空兴趣点[10]、立体兴趣点[11]、运动历史图像[12]、光流直方图(HOF)[13] 等局部描述符，通过视觉词袋[14]、Fisher Vector[15] 等特征融合方法，用 KNN、SVM 等传统分类器进行分类。在 2015 年以前，iDT[16] 是行为识别领域精度最高的方法。该方法通过提升的密集轨迹方法对相机运动进行估计，使用行人检测消除干扰信息，再基于光流直方图和光流梯度直方图等描述子进行 SVM 分类。iDT 方法识别效果优良、鲁棒性好，但人工特征提取流程复杂且特征不够全面。随着深度神经网络的不断发展，基于深度学习的方法在精度和计算成本上都超越了传统方法。

目前，基于深度学习的行为识别方法有双流网络、循环神经网络、3D 卷积等。

视频理解除了空间信息之外还需要运动信息，双流网络使用两个并行的卷积神经网络，分别独立进行特征提取，主流的双流方法有 TSN[17]、Convolutional Two-Stream[18]、Flownet[19] 等。在经典的 Two-steam[20] 方法中，一个网络处理单帧的图像，提取环境、视频中的物体等空间信息，另一个网络使用光流图做输入，提取动作的动态特征。考虑到光流是一种手工设计的特征，双流方法通常都无法实现端到端的学习。另外，随着行为识别领域数据集规模的不断扩大，由光流图计算带来的巨大计算成本和存储空间消耗等问题使得基于光流的双流卷积神经网络不再适用于大规模数据集的训练和实时

部署。

LSTM[21]是循环神经网络中一种，该网络用于解决某些动作的长依赖问题。文献[22]研究了同时使用卷积网络和循环神经网络的 CNN-LSTM 网络结构在行为识别任务中的表现，发现需要对视频进行预分段，LSTM 才能提取到较为明确的时间信息。文献[23]探索了多种 LSTM 网络在行为识别任务中的应用效果，发现相比于行为识别，LSTM 更适合于动作定位任务。在视频行为识别中，很大一部分动作只需要空间特征就能够识别，但 LSTM 网络只能对短时的时间信息进行特征提取，无法很好地处理空间信息。因此，该类方法已逐渐被 3D 卷积等主流方法取代。

视频行为识别中，主流的 3D 卷积方法有 C3D[24]、I3D[25]、P3D[26]等。文献[27]将经典的残差神经网络 ResNet 由 2D 拓展为 3D，并在各种视频数据集中探索了从较浅到深的 3D ResNet 体系结构，结果发现在大规模数据集上，较深的 3D 残差神经网络能够取得更好的效果。然而，视频信息在时空维度具有完全不同的分布方式和信息量，经典的 3D 卷积方法在时空维度并没有对此进行区分，计算了过多的冗余信息，由此带来了过高的计算代价以及部署成本。

文献[8]提出了一种受生物机制启发的行为识别模型，通过分解架构分别处理空间信息和时间信息。在人类视觉中，空间语义（颜色、纹理、光照等）信息变化较慢，可使用较低的帧率。相比之下，大部分动作（拍手、挥手、摇晃、走路、跳跃等）比空间语义信息变化速度快得多，因此使用更高的帧率来进行有效建模。但是，该方法只改变了两个路径输入视频帧的数量。对单个视频帧没有进行更细致的处理，在空间流也未添加更多的动作信息予以辅助。

当前，已经存在很多基于 3D 卷积和双路径网络架构的行为识别方法，但效果均不理想，这主要是由于对于行为识别任务，视频中的信息较为冗余，对任务做出实际贡献的视频帧和含有动作信息的特征通道在视频中的分布十分稀疏。因此，如何找出含有关键信息的视频帧和特征通道亟待解决。

2 MRTP 方法

本文设计了一个时间与动作感知的双路径行为识别方法 MRTP，网络结构见图 1。模型使用双路径结构，以视频包为输入，在时间维度上以步长 1 为滑动窗口，可得到视频中顺序排列的连续 n 帧图像。

每个视频以 2 s 长度截取视频包，对于视频包中的 64 帧图像再进行采样。T 为每次采样的视频帧数，在高帧率动作路径设置 $T=32$，低帧率空间路径设置 $T=4$。低帧率空间路径所取视频帧的位置由高帧率动作路径的时间注意力模块生成的 α 和 β 决定，α 和 β 为时间注意力筛选出的权重最大两帧图像对应的坐标。

高帧率动作路径采样的图像数量较多但通道数较少，低帧率空间路径采样的图像数量较少但通道数较多。设高帧率动作路径输入的图像数为低帧率空间路径的 p 倍，高帧率动作路径特征的通道数为低帧率空间路径的 q 倍，在 UCF-101 数据集和 AVA 数据集上，$p=8$，$q=1/16$。

Res1～Res4 是 ResNet3D 的残差结构。使用

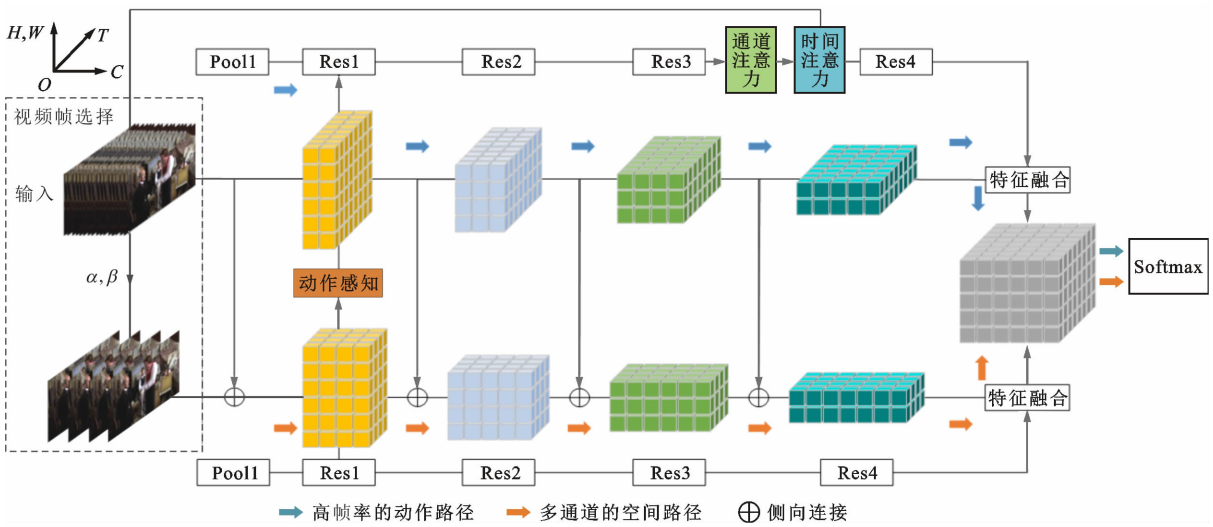


图 1 MRTP 行为识别方法网络架构

Fig. 1 Network architecture of MRTP action recognition method

Kinetics 400 和 Kinetics 600 上预训练的 ResNet3D 50 和 ResNet3D 101 作为特征提取的骨干网络。

通道注意力模块用于衡量动作路径各个特征通道的重要性并进行加权。时间注意力模块在通道注意力模块筛选出的通道权重基础上衡量各个视频帧的重要性，将 α 和 β 输入到低帧率空间路径作为图像提取的位置坐标依据。动作感知模块基于相邻两帧的特征差分矩阵衡量前后两个视频帧的特征变化，并对通道赋予权重。

在卷积网络的 Pool1、Res1、Res2、Res3 之后建立侧向连接，将动作路径的特征通过重构之后传递到空间路径。

特征融合部分将高帧率动作路径和低帧率空间路径的特征连接起来。

Softmax 函数将融合后的特征向量转换为类别概率向量，并选取其中的最大值所对应的类别作为输出结果。

2.1　高帧率动作路径特征提取

2.1.1　*长时间跨度动作特征*　在由图像序列组成的视频数据中，动态信息被定义为帧间图像的像素运动，即光流。然而，光流需要时间的变化不引起目标位置的剧烈变化，因此光流矢量只能在帧间位移较小的前提下使用。在需要长时间跨度动作特征提取的情况下，光流作为动态信息的一种表示，并不能提取出所需的动作信息表征。因此，本文引入高帧率采样的动作路径，该路径输入 RGB 视频帧，在本文实验的两个数据集上将帧率变为原来的 p 倍。同时，为了降低模型的计算量，使该路径更加聚焦于动态信息，本文将动作路径的通道数量变为原来的 q 倍，在保证了模型轻量化的同时实现了动态信息的提取。相比于基于光流的动态信息，本文通过使用 RGB 视频帧输入实现了端到端的训练和部署，并且特征的提取不再受光流的场景固定和小范围时间跨度的约束。

2.1.2　*通道注意力机制*　由于输入特征向量在通道维度有较大差异，有的通道对识别任务有较大贡献，但部分通道贡献较小，所以在 3D 卷积中引入通道注意力机制。将提取特征向量作为输入，通过计算通道权重对通道加权。

设输入特征向量的维度用数组 X 表示，$X=[N,C,\omega T,W,H]$，其中：N 为输入的视频数；C 为通道数量；ω 为整个视频中所取的片段数，即进行 3D 卷积的次数，若视频长度在 2 s 以内，则 $\omega=1$；W 和 H 为特征的宽和高。首先，在时间维度对特征进行融合

$$\boldsymbol{u}_{\mathrm{C}}=\boldsymbol{v}_{\mathrm{C}}*\boldsymbol{x}=\sum_{i=1}^{\omega T}\boldsymbol{v}_{\mathrm{C}}^{i}*\boldsymbol{x}^{i} \tag{1}$$

式中：$*$ 表示卷积操作；$\boldsymbol{u}_{\mathrm{C}}$ 为时间维度的卷积结果；$\boldsymbol{v}_{\mathrm{C}}$ 为卷积核；$\boldsymbol{v}_{\mathrm{C}}^{i}$ 为第 i 帧图像对应的卷积核；$\boldsymbol{x}^{i}$ 为第 i 帧图片对应的特征向量。通过第 1 步卷积操作，特征向量维度变化为 $X=[N,C,1,W,H]$。

然后，在空间维度通过池化融合特征

$$\boldsymbol{z}_{\mathrm{C}}=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{u}_{\mathrm{C}}(i,j) \tag{2}$$

式中 $\boldsymbol{z}_{\mathrm{C}}$ 为池化操作的结果。通过在特征的宽和高进行池化，特征向量的维度变为 $X=[N,C,1,1,1]$。

最后，计算出每个通道的权重向量

$$\boldsymbol{a}=\mathrm{Sigmoid}(\boldsymbol{Y}_2\mathrm{ReLU}(\boldsymbol{Y}_1\boldsymbol{z}_{\mathrm{C}})) \tag{3}$$

式中：$\boldsymbol{a}$ 为通道注意力计算出的权重向量；$\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$ 为权重参数，在训练中得到；Sigmoid 为 S 型激活函数；ReLU 为线性激活函数。

2.1.3　*时间注意力机制*　由于每帧图像的重要性不同，所以对于通道加权后的特征向量，选取其中权值最大的通道特征作为时间注意力机制的输入并计算权重，从而对视频帧加权。

首先，利用输入的通道权重对通道数据进行筛选

$$\boldsymbol{u}_{\mathrm{T}}=\boldsymbol{x}[N,\boldsymbol{a}_{\max},\omega T,W,H] \tag{4}$$

式中：$\boldsymbol{x}$ 为输入特征向量；$\boldsymbol{a}_{\max}$ 为上一步通道注意力机制中提取出的权重最大值对应的通道坐标；$\boldsymbol{u}_{\mathrm{T}}$ 为通道注意力提取出的权重最大通道对应的特征向量。通过第 1 步提取操作，特征向量维度变化为 $X=[N,1,\omega T,W,H]$。

然后，在空间维度通过池化融合特征

$$\boldsymbol{z}_{\mathrm{T}}=\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{u}_{\mathrm{T}}(i,j) \tag{5}$$

式中 $\boldsymbol{z}_{\mathrm{T}}$ 为池化操作的输出特征。通过在特征的宽和高进行池化，特征向量的维度变化为 $X=[N,1,\omega T,1,1]$。

最后，计算出每个视频帧的权重向量

$$\boldsymbol{s}=\mathrm{Sigmoid}(\boldsymbol{W}_2\mathrm{ReLU}(\boldsymbol{W}_1\boldsymbol{z}_{\mathrm{T}})) \tag{6}$$

式中：$\boldsymbol{s}$ 为时间注意力计算出的权重向量；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 为权重参数，在训练中得到。

2.2　低帧率空间路径特征提取

2.2.1　*视频帧按权重采样*　空间路径采样视频帧的数量只有动作路径的 $1/p$，在空间路径使用均匀采样会因为位置不准确导致无法提取出足够的信息。因此，MRTP 方法采用动作路径生成的权重对空间路径进行非均匀采样指导，流程如图 2 所示。

动作路径中的通道注意力和时间注意力模块生成了视频帧权重。基于该权重，在空间路径按权值从大到小，以 2 帧/s 的处理速度在视频对应位置采样图像。假设时间注意力计算出的权重 $\boldsymbol{s}$ 中最大的两个值为 s_α 和 s_β，则在视频中按 α 和 β 所在位置抽取图像。相比于现有模型均匀抽取的方法，这种采样方法能够提取到信息量更多、对识别贡献更大的视频帧。

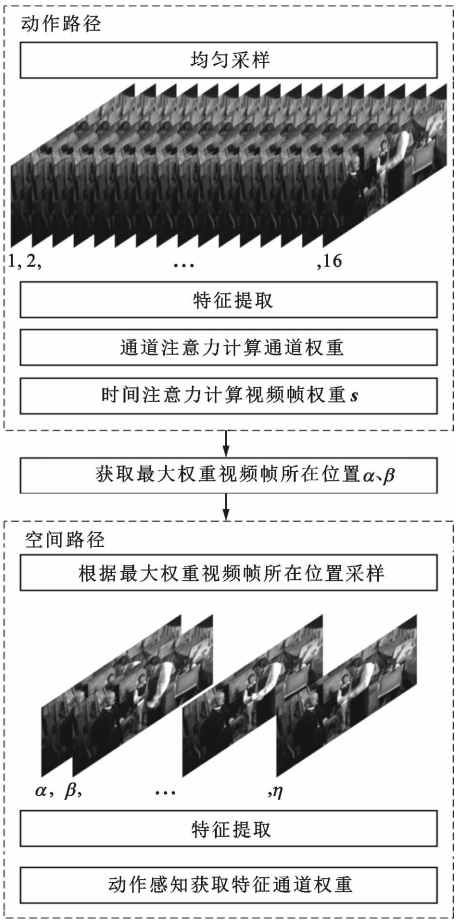


图 2 空间路径视频帧按动作路径时间注意力权重进行非均匀采样示意

Fig. 2 Non-uniform sampling in spatial path according to time attention weight in motion path

2.2.2 动作空间特征提取 空间特征主要描述动作中涉及到的物体外观和场景配置。为了提取视频帧中细节的空间信息，本文使用低帧率空间路径，一次卷积中只使用 4 帧图像。预处理随机裁剪将图像归一化为 224×224 像素，在训练出的 ResNet-3D 网络模型中，Res4 的特征通道数达到了 2 048。更多的特征通道能够让该路径提取到颜色、纹理、背景等细节的空间信息。

2.3 动作感知

为了替代以光流为基础的像素级动作表示方式，并将时空特征结合起来，本文在低帧率空间路径使用了动作感知模块，从特征通道来进行动作表征和激励。该模块通过衡量前后两个视频帧的特征变化，赋予视频帧中动作信息对应的特征通道更大的激励权重，以此来增强网络对动作的感知能力。动作感知模块的计算流程如图 3 所示。

设输入特征为 $\boldsymbol{X}$，$\boldsymbol{X}$ 的特征维度即为 $X=[N,C,\omega T,W,H]$，此处 $\boldsymbol{X}$ 为一次卷积获得的特征，即 $\omega=1$，可得 $X=[N,C,T,W,H]$。首先，使用一个 3D 卷积层来降低通道数以提高计算效率

$$\boldsymbol{X}^k = \mathrm{conv}_{3D}(\boldsymbol{X}) \tag{7}$$

式中：$\boldsymbol{X}^k$ 表示通道减少后的特征，$\boldsymbol{X}^k$ 特征维度为 $[N,C/k,T,W,H]$，$k=16$ 是减少的比率；conv_{3D} 表示使用尺寸为 1×1×1 的卷积核对通道维度进行降维操作。

对于运动特征向量，使用前后两帧图像对应的特征 $\boldsymbol{X}^k(t+1)$ 和 $\boldsymbol{X}^k(t)$ 之间的差来表示运动信息

$$\boldsymbol{P}(t) = \mathrm{conv}_{\mathrm{shift}}(\boldsymbol{X}^k(t+1)) - \boldsymbol{X}^k(t) \tag{8}$$

式中：$\boldsymbol{P}(t)$ 是时间 t 时的动作特征向量，特征维度为 $[N,C/k,1,W,H]$，$1\leqslant t\leqslant T-1$；$\mathrm{conv}_{\mathrm{shift}}$ 是一个 3×32 通道卷积层，对每个通道进行转换。

假设 T 时刻动作已经结束，即 T 时刻已经没有动作特征，令 $\boldsymbol{P}(T)$ 为 0 特征向量。在计算出每个时刻的 $\boldsymbol{P}(t)$ 之后，构造出整个 T 帧序列的动作矩阵 $\boldsymbol{P}$。通过全局平均池化层激发对动作敏感的通道

$$\boldsymbol{P}^l = \mathrm{pool}(\boldsymbol{P}) \tag{9}$$

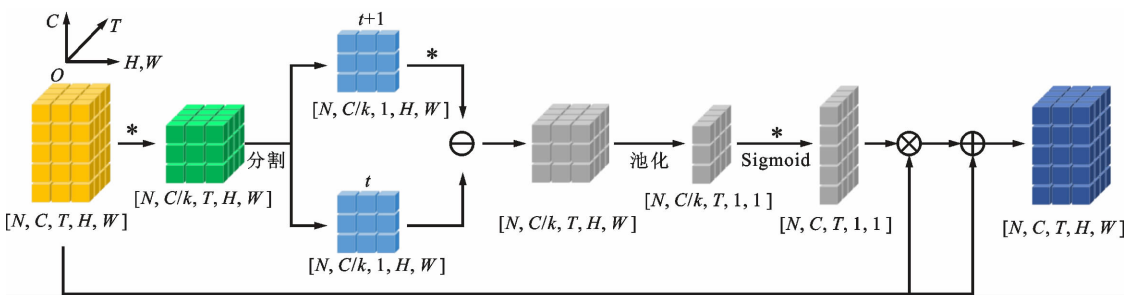


图 3 动作感知结构

Fig. 3 Motion perception structure

式中 $\boldsymbol{P}^l$ 特征维度为[N,C/k,T,W,H]。使用 3D 卷积层将动作特征的通道维度 C/k 扩展到原始通道维度 C，再利用 Sigmoid 函数得到动作感知权值

$$\boldsymbol{E} = 2\text{Sigmoid}(\text{conv}_{3\text{D}}(\boldsymbol{P}^l)) - 1 \tag{10}$$

至此，得到了特征向量中各通道的动作相关性权重 $\boldsymbol{E}$。为了不影响原低帧率动作路径的空间特征信息，借鉴 ResNet 中残差连接的方法，在增强动作信息的同时保留原有的空间信息

$$\boldsymbol{X}^{\text{R}} = \boldsymbol{X} + \boldsymbol{X} \odot \boldsymbol{E} \tag{11}$$

式中：$\boldsymbol{X}^{\text{R}}$ 是该模块的输出；⊙表示按通道的乘法。

3　实　验

3.1　实验设置

3.1.1　损失函数　在训练过程当中，对于同一输入有多个动作共存的情况，Sigmoid 函数计算公式为

$$S(x) = \frac{1}{1+\text{e}^{-x}} \tag{12}$$

由于经过 Sigmoid 网络层后的输出为[0,1]内的概率值，因此本文选择二分类交叉熵损失函数进行训练，即对每一类动作都进行二分类判别。在判别时设定概率阈值为 0.8，当大于该阈值时认为判别有效，即视频中包含该类动作，从而避免多分类的类别互斥情况，损失函数计算公式为

$$L = -\frac{1}{N}\sum_{i=1}^{N'}\sum_{j=1}^{C'}[\boldsymbol{y}_j^{(i)}\log(\hat{\boldsymbol{y}}_j^{(i)}) + (1-\boldsymbol{y}_j^{(i)})\log(1-\hat{\boldsymbol{y}}_j^{(i)})] \tag{13}$$

式中：L 为损失值；N' 为样本数量；C' 为类别数量；$\boldsymbol{y}_j^{(i)}$ 为第 j 个类别中的第 i 个样本的真实值；$\hat{\boldsymbol{y}}_j^{(i)}$ 为第 j 个类别中的第 i 个样本的预测值。

3.1.2　训练参数　本文实验使用深度学习框架 Pytorch 实现，训练使用 SGD 优化器，学习率调整策略为 StepLR，基于 epoch 训练次数进行学习率调整，即每到给定的 epoch 数时，学习率都改变为初始学习率的指定倍数。初始学习率设置为 0.05，指定当 epoch 数为 10、15、20 时，学习率分别设置为初始学习率的 0.1、0.01、0.001 倍，权重衰减设置为 1×10^{-7}，Dropout rate 设置为 0.5。AVA 数据集训练样本庞大，刚开始采用较大的学习率可能会带来模型不稳定。为了防止出现提前过拟合的现象和保持分布的平稳，本文在训练过程中还加入了学习率预热策略，在 epoch 数小于 5 时，使用0.000 125的学习率进行训练，当模型具备了一定的先验知识，再使用预先设置的学习率，这样可以避免初期训练时错过最优点导致损失振荡，从而加快模型的收敛速度。

3.2　数据集

本文使用两个数据集评估 MRTP 的性能。其中，UCF101 是行为识别领域的经典数据集，AVA2.2 是目前最具挑战性的大规模数据集。在 UCF101 和 AVA2.2 上，分别使用三折交叉验证准确率和平均精度(mAP)作为评价指标，与经典方法以及近期方法进行了对比，并单独验证了 MRTP 的有效性。

3.2.1　UCF101　UCF101[28]是一个由佛罗里达大学创建的动作识别数据集，收集自 YouTube。UCF101 拥有来自 101 个动作类别的13 320个视频，在摄像机运动、外观、姿态、比例、视角、背景、照明条件等方面存在很大的差异。101 个动作类别中的视频被分成 25 组，每组可以包含一个动作的 4～7 个视频。同一组视频可能有一些共同特点，比如相似的背景或类别等。数据集包括人与物体交互、单纯的肢体动作、人与人交互、演奏乐器、体育运动共 5 大类动作。

3.2.2　AVA　AVA 数据集[29]来自谷歌实验室，包含 430 个视频，其中，235 个用于训练，64 个用于验证，131 个用于测试。每个视频有 15 min 的注释时间，间隔为 1 s。尽管很多数据集采用了图像分类的标注机制，即数据的每一个视频片段分配一个标签，但是仍然缺少包含不同动作的多人复杂场景数据集。与其他动作数据集相比，AVA 具备每个动作标签都与人更加相关的关键特征。在同一场景中执行不同动作的多人具有不同的标签。AVA 的数据源来自不同类型和国家的电影，覆盖大多数的人类行为并且十分贴近实际部署情况。相比于 AVA2.1，AVA2.2 数据源没有变化，但在标签文件中添加了 2.5%的缺失动作标签。

相比于传统的 UCF101 和 HMDB51 等数据集，AVA 数据集十分具有挑战性，该数据集的数据量是传统数据集的数 10 倍，场景切换十分频繁，除了相机运动带来的场景连续变化，还出现了电影镜头切换带来的场景突变。相比于主流的 Kinetics 和 Youtube-8M 等数据集，AVA 数据集使用了多人标注，在更加贴近真实场景的同时，增加了对人的检测和跟踪，人数增多和遮挡问题也造成了包含单个动作的源数据大幅减少。因此，该数据集识别难度远超现有的其他主流数据集。在此之前，文献[8]训练的模型达到了 27.1%的 mAP 精度(由文献[30]进行复现和评估)，是该数据集上的最高精度。

3.3 评价指标

3.3.1 准确率 准确率为分类正确的样本数占总样本的比例,公式为

$$A=\frac{1}{m}\sum_{i=1}^{M}I(f(x_i)) \tag{14}$$

式中:A 为准确率;m 为总样本数;$f(x_i)$为第 i 个样本 x_i 的预测分类结果;y_i 为 x_i 的实际分类结果;I 为判别函数,当样本 x_i 的分类结果与实际结果 y_i 相同时,$I(f(x_i)=y_i)=1$,否则 $I(f(x_i)\neq y_i)=0$。

在 UCF-101 中使用三折交叉验证准确率作为评价指标。将数据集平均分成 3 份,使用其中 1 份作为测试数据,其余作为训练数据。在 3 份数据上重复进行这个训练测试过程,取最后的测试准确率平均值作为结果。

3.3.2 mAP AP 是某一类 P-R 曲线下的面积,mAP 则是所有类别 P-R 曲线下面积的平均值。P-R 曲线是以查全率为横坐标、查准率为纵坐标构成的曲线。查全率公式为

$$R=\frac{T'}{T'+N'} \tag{15}$$

式中:R 为查全率;T'为真阳性数,表示交并比大于 0.5 的检测框数;N'为假阴性数,表示交并比小于 0.5 的检测框数。查准率公式为

$$P=\frac{T'}{T'+F} \tag{16}$$

式中:P 为查准率;F 为假阳性数,表示漏检的真实检测框的数量。

AVA 数据集中存在同一场景多人同时执行动作的情况,因此需要目标检测来区分每个人对应的动作,使用 mAP 来衡量实验结果。

3.4 实验结果

3.4.1 UCF101 实验结果 使用 Kinetics-400 数据集进行预训练,在预训练模型的基础上对 UCF-101 数据集的行为识别数据进行训练建模,对 UCF-101 的 3 个 split 进行测试,与同样使用 3D 卷积的 C3D[24] 方法和同样使用了双路径结构的 TSN[17]、Two-stream I3D[7] 以及近期的 I3D-LSTM[31]、TesNet[32] 进行了准确率的对比,结果如表 1 所示。可以看出,相比于主流的行为识别方法,本文在同样的数据集上取得了更高的测试精度。

表 1 UCF101 数据集上不同方法的准确率对比

Table 1 Comparison of model accuracy for different methods on UCF101 data set

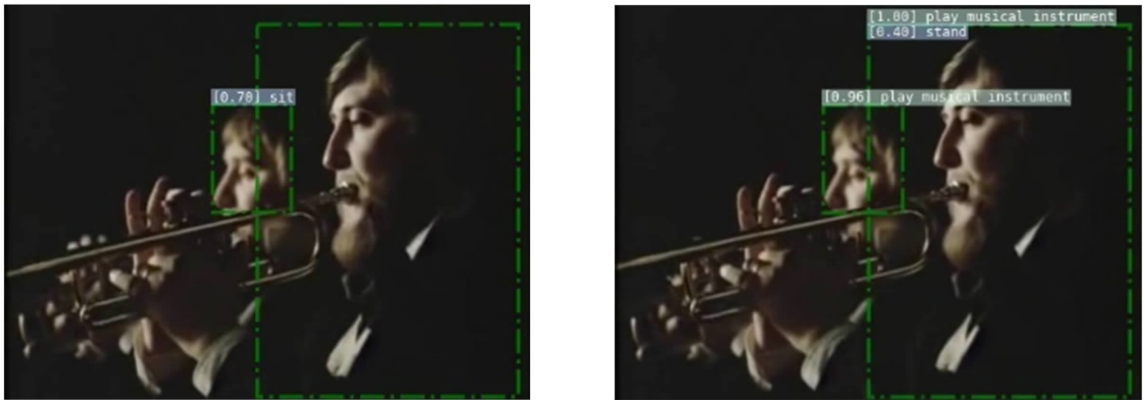
方法	是否使用光流	骨干网络	预训练模型	Top-1 准确率/%	Top-5 准确率/%
C3D(2014)	是	ResNext-101	Sports-1M	82.3	95.9
TSN-RGB(2016)	否	ResNet50	ImageNet	83.0	96.7
Two-stream I3D(2017)	是	Inception-v1	ImageNet	90.6	
TSN-Optical Flow(2016)	是	ResNet-50	Kinects	91.7	99.2
I3D-LSTM(2019)	否	LSTM	Kinetics	95.1	
TesNet(2020)	是	Inception-ResNet-V2	ImageNet	95.2	
MRTP(2021)	否	ResNet3D	Kinects	**95.6**	**99.5**

注:表中加粗数据为最优值。

3.4.2 AVA2.2 实验结果 同一视频片段识别结果对比示例如图 4 所示,该视频片段真实的动作标签为“站立(stand)”和“演奏乐器(play musical instrument)”。基础模型使用了 2 帧/s 的固定帧率对视频进行采样,未加入本文提出的 MRTP 方法,同样使用 ResNet3D 作为骨干网络。在使用基础模型和本文提出的 MRTP 方法对相同输入进行识别时,基础模型无法正确地识别出动作类别,识别出的结果为“坐(sit)”,而本文提出的 MRTP 方法在同样的输入数据下相比基础模型有更准确的识别结果。

在 Kinetics-400 和 Kinetics-600 上进行预训练,得到含有低层基础特征的预训练模型,基于预训练模型对 AVA2.2 的数据进行训练建模。在测试集上计算交并比阈值为 0.5 时的 mAP 精度,ARCN[33]、I3D w/RPN[34]、I3D Tx HighRes 方法在 AVA2.1 上进行了测试,AVA 数据集上的 mAP 精度结果如表 2 所示。可以看出,相比于 ARCN[33]、I3D w/RPN[34]、I3D Tx HighRes[35]、D3D[36] 和 X3D[30] 等行为识别方法,MRTP 取得了更高的测试精度。在网络深度相同的情况下,MRTP 超过了之前效果最好的 SlowFast 方法,在加深骨干网络到 101 层之后,MRTP 达到了 28.0%的 mAP 精度,刷新了目前 AVA2.2 数据集上最高的 mAP 精度。

3.4.3 ResNet3D 骨干网络实验结果 为了证明本文 MRTP 方法的有效性,固定了骨干网络和预训练模型,在两个数据集上对比了添加 MRTP 方法前



(a)基础模型识别结果　　(b)MRTP 识别结果

图 4　同一视频片段识别结果对比示例

Fig. 4　Comparison example of recognition results

表 2　AVA 数据集上不同方法的 mAP 对比

Table 2　Comparison of model mAP accuracy for different methods on the AVA data set

方法	是否使用光流	骨干网络	预训练模型	平均精度/%
ARCN(2018)	是	S3D-G		17.4
I3D w/RPN(2018)	否	ResNet50	ImageNet	21.9
I3D Tx HighRes(2019)	否	I3D Transformer	Kinetics-400	27.6
D3D(2018)	否	ResNet RPN	Kinetics-400	23.0
X3D-M(2020)	否	ResNeXt-101	Kinetics-600	23.2
SlowFast 4x16 R50(2019)	否	ResNet3D-50	Kinetics-400	21.9
SlowFast 8x8 R101＋NL(2019)	否	ResNet3D-101	Kinetics-600	27.1
X3D-XL(2020)	否	ResNeXt-101	Kinetics-600	27.4
SlowFast 16x8 R101＋NL(2019)	否	ResNet3D-101	Kinetics-600	27.5
MRTP(2021)	否	ResNet3D-50	Kinetics-400	**25.2**
MRTP(2021)	否	ResNet3D-101	Kinetics-600	**28.0**

注：表中加粗数据为最优值。

后的评价指标，结果如表 3 所示。可以看出，相比于基础模型，添加了 MRTP 方法后在不同的数据集和网络深度都能够实现精度的提升。

加入 MRTP 方法前后，部分类别 mAP 精度对比见表 4。可以看出，在基础模型中加入本文提出的 MRTP 方法后，AVA 数据集中大部分行为类别的准确率都有了一定程度的提升，特别是“演奏乐器(play musical instrument)”，“射击(shoot)”以及“游泳(swim)”这 3 类动作，更是取得了 10%以上的提升。原因在于本文使用的时间注意力和动作感知方法都是聚焦于动作的动态信息。这 3 类动作都是在视频画面中动作变化相对较小的。在所提取的特

表 3　加入 MRTP 方法前后的对比结果

Table 3　Comparison of the results before and after adding MRTP method

方法	骨干网络	预训练模型	UCF-101		AVA2.2
			Top-1 准确率/%	Top-5 准确率/%	平均精度/%
基础模型	ResNet3D-50	Kinetics-400	91.3	99.2	22.0
MRTP			**93.5**	**99.2**	**25.2**
基础模型	ResNet3D-101	Kinetics-600	94.3	99.5	27.1
MRTP			**95.6**	**99.5**	**28.0**

注：表中加粗数据为最优值。

征中,这类变化较小的动作信息容易被场景、光线、角度变化所干扰,而 MRTP 在时间维度使用时间注意力聚焦于含有动作变化的视频帧,在通道维度使用特征差分的动作感知聚焦于含有动作信息的通道。这样就使得模型所获取的动态信息大多来自于动作本身,从而在这些动态信息不明显的动作类别上实现 mAP 精度的提升。

表 4 AVA 数据集加入 MRTP 方法前后的部分类别 mAP 精度对比

Table 4 Comparison of accuracy of some categories before and after adding the MRTP method to the AVA data set 单位:%

行为类别	基础模型准确率	MRTP 准确率	准确率提升值
打电话(answer phone)	66.73	72.11	5.38
跳舞(dance)	50.82	56.87	6.06
饮用(drink)	24.86	29.55	4.69
握手(hand shake)	12.76	19.76	7.00
演奏乐器(play musical instrument)	41.08	51.68	**10.60**
射击(shoot)	9.02	22.70	**13.67**
游泳(swim)	38.28	57.57	**19.29**
打开某物(open)	18.64	16.75	−1.89

注:表中加粗数据为大于 10%的准确率提升值。

3.4.4 复杂度分析 各方法训练出的模型复杂度对比见表 5。可以看出:本文提出的 MRTP 方法在使用 ResNet3D-50 作为骨干网络时的参数量小于同样使用 3D 卷积网络的 I3D-NL 方法[37]的,甚至小于使用 2D 卷积网络的 TSN 方法的;同样使用 RTX 3090 显卡进行模型测试,输入同一个分辨率为 640 × 480 像素的测试视频,MRTP 达到了 110.24 帧/s 的处理速度,在所有方法中是最优的,虽然使用 ResNet3D-101 作为骨干网络时模型参数量较大,但是处理速度依然远超使用了光流输入的 TSN 方法[17]的,也高于使用伪 3D 卷积的 R2+1D[38]方法的。本文方法使用 RGB 视频作为输入,极大地减少了由于计算光流图带来的时间和计算成本,并且通过在动作路径将特征通道数量减少,使得在动作路径增加的输入视频帧没有带来更大的计算消耗。

表 5 不同方法的模型复杂度对比

Table 5 Comparison of model complexity for different methods

方法	骨干网络	模型参数量/10^{12}	处理速度/(帧·s^{-1})
R2+1D	ResNet-50	53.950	64.06
TSN	ResNet-101	43.320	7.66
I3D-NL	ResNet-101	61.780	101.94
MRTP	ResNet3D-50	**34.566**	**110.24**
MRTP	ResNet3D-101	62.827	75.62

注:表中加粗数据为最优值。

4 结 论

针对时空信息分布不均衡以及对长时间跨度信息表征获取难的问题,本文提出了一种时间-动作感知的多尺度时间序列实时行为识别方法 MRTP。本文得出的主要结论如下。

(1)提出的网络使用双路径结构,在不同的时间分辨率上对视频进行特征提取,相比于只使用固定帧率的网络,对长时动作能够更好地聚焦于时序信息。

(2)在低帧率空间路径中,使用基于特征差分的动作感知寻找并加强通道动作特征,将变化明显的特征通道作为动作的表征;在高帧率动作路径中加入通道注意力和时间注意力加强关键特征,细化了各个视频帧的重要性度量。

(3)低帧率空间路径基于动作路径中的时间注意力生成的视频帧权重对输入视频进行采样,相比于现有方法的均匀采样,能够提取到识别贡献更大的视频帧;在高帧率动作路径中,基于空间路径动作感知的权重进行通道筛选,保留了动作信息丰富的特征通道。

(4)本文提出的 MRTP 方法仅使用 RGB 帧作为输入,通过衡量帧权重,在时序维度上获得了更好的依赖,通过动作感知寻找并加强了通道维度动作特征表征。两个路径的信息交互和指导使得整个网络更加聚焦于动作信息在时间和通道所处的位置。本文方法在公共数据集上表现出良好的识别性能,在 AVA2.2 数据集上达到了 28%的 mAP 精度,刷新了 AVA2.2 数据集目前最高的 mAP 精度。不同环境的实验结果也表明了 MRTP 良好的鲁棒性。

(5)在未来的工作中,将从时序特征出发,通过特征差分提取更为有效和显式的时序信息表征,并继续探索双路径网络并行分支互相交互的可能性。

参考文献:

[1] 朱煜,赵江坤,王逸宁,等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报,2016,42(6):848-

857.
ZHU Yu, ZHAO Jiangkun, WANG Yining, et al. A review of human action recognition based on deep learning [J]. Acta Automatica Sinica, 2016, 42(6): 848-857.

[2] 罗会兰，王婵娟，卢飞. 视频行为识别综述 [J]. 通信学报，2018，39(6)：169-180.
LUO Huilan, WANG Chanjuan, LU Fei. Survey of video behavior recognition [J]. Journal on Communications, 2018, 39(6): 169-180.

[3] ZHU Yi, LI Xinyu, LIU Chunhui, et al. A comprehensive study of deep video action recognition [EB/OL]. [2021-04-01]. https://arxiv.org/abs/2012.06567.

[4] VAROL G, LAPTEV I, SCHMID C. Long-term temporal convolutions for action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1510-1517.

[5] CHEN Yuehai, YANG Jing, ZHANG Kun, et al. A feature-cascaded correntropy LSTM for tourists prediction [J]. IEEE Access, 2021, 9: 32810-32822.

[6] LI Zhenyang, GAVRILYUK K, GAVVES E, et al. VideoLSTM convolves, attends and flows for action recognition [J]. Computer Vision and Image Understanding, 2018, 166: 41-50.

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

[8] FEICHTENHOFER C, FAN Haoqi, MALIK J, et al. SlowFast networks for video recognition [C]// Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 6201-6210.

[9] ZHANG Shiwen, GUO Sheng, HUANG Weilin, et al. V4D: 4D convolutional neural networks for video-level representation learning [C/OL]// Proceedings of the 2020 International Conference on Learning Representations. London, UK: ICLR, 2020 [2021-04-01]. https://arxiv.org/pdf/2002.07442.pdf.

[10] LAPTEV I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(2): 107-123.

[11] DOLLAR P, RABAUD V, COTTRELL G, et al. Behavior recognition via sparse spatio-temporal features [C]// Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. Piscataway, NJ, USA: IEEE, 2005: 65-72.

[12] BOBICK A F, DAVIS J W. The recognition of human movement using temporal templates [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-267.

[13] LAPTEV I, MARSZALEK M, SCHMID C, et al. Learning realistic human actions from movies [C]// Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2008: 1-8.

[14] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [C]// Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2006: 2169-2178.

[15] PERRONNIN F, DANCE C. Fisher kernels on visual vocabularies for image categorization [C]// Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2007: 1-8.

[16] WANG Heng, SCHMID C. Action recognition with improved trajectories [C]// Proceedings of the 2013 IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2013: 3551-3558.

[17] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: towards good practices for deep action recognition [C]// Proceedings of the 2016 European Conference on Computer Vision. Cham, Germany: Springer, 2016: 20-36.

[18] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 1933-1941.

[19] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 1647-1655.

[20] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS, 2014: 568-576.

[21] HOCHREITER S, SCHMIDHUBER J. Long short-

term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[22] NG J Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification [C]// Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2015: 4694-4702.

[23] MA C Y, CHEN M H, KIRA Z, et al. TS-LSTM and temporal-inception: exploiting spatiotemporal dynamics for activity recognition [J]. Signal Processing: Image Communication, 2019, 71: 76-87.

[24] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]// Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2015: 4489-4497.

[25] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset [EB/OL]. [2021-04-01]. https://arxiv.org/abs/1705.06950.

[26] QIU Zhaofan, YAO Ting, MEI Tao. Learning spatio-temporal representation with pseudo-3D residual networks [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 5534-5542.

[27] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 6546-6555.

[28] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. [2021-04-01]. https://arxiv.org/abs/1212.0402.

[29] GU Chunhui, SUN Chen, ROSS D A, et al. AVA: a video dataset of spatio-temporally localized atomic visual actions [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 6047-6056.

[30] FEICHTENHOFER C. X3D: expanding architectures for efficient video recognition [C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 200-210.

[31] WANG Xianyuan, MIAO Zhenjiang, ZHANG Ruyi, et al. I3D-LSTM: a new model for human action recognition [C]// Proceedings of the IOP Conference Series: Materials Science and Engineering. London, UK: IOP, 2019: 032035.

[32] HUANG Guoxi, BORS A G. Learning spatio-temporal representations with temporal squeeze pooling [C]// Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ, USA: IEEE, 2020: 2103-2107.

[33] SUN Chen, SHRIVASTAVA A, VONDRICK C, et al. Actor-centric relation network [C]// Proceedings of the 2018 European Conference on Computer Vision. Cham, Germany: Springer, 2018: 335-351.

[34] GIRDHAR R, CARREIRA J, DOERSCH C, et al. A better baseline for AVA [EB/OL]. [2021-04-01]. https://arxiv.org/abs/1807.10066.

[35] GIRDHAR R, JOÃO CARREIRA J, DOERSCH C, et al. Video action transformer network [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2019: 244-253.

[36] STROUD J C, ROSS D A, SUN Chen, et al. D3D: distilled 3D networks for video action recognition [C]// Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway, NJ, USA: IEEE, 2020: 614-623.

[37] WANG Xiaolong, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7794-7803.

[38] TRAN D, WANG Heng, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 6450-6459.

(编辑 陶晴)