计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2022/16(09)-2061-07
doi: 10.3778/j.issn.1673-9418.2012119

时空特征金字塔模块下的视频行为识别

龚苏明,陈　莹+
江南大学 轻工过程先进控制教育部重点实验室,江苏 无锡 214122
+通信作者 E-mail: chenying@jiangnan.edu.cn

摘　要:目前用于视频行为识别的主流2D卷积神经网络方法无法提取输入帧之间的相关信息,导致网络无法获得输入帧间的时空特征信息进而难以提升识别精度。针对目前主流方法存在的问题,提出了通用的时空特征金字塔模块(STFPM)。STFPM由特征金字塔和空洞卷积金字塔两部分组成,并能直接嵌入到现有的2D卷积神经网络中构成新的行为识别网络——时空特征金字塔网络(STFP-Net)。针对多帧图像输入,STFP-Net首先提取每帧输入的单独空域特征信息,并将这些特征信息记为原始特征;然后,所设计的STFPM利用矩阵转换操作对原始特征构建特征金字塔;其次,利用空洞卷积金字塔对构建的原始特征金字塔提取具有时空关联性的时序特征;接着,将原始特征与时序特征进行加权融合并传递给后续深层网络;最后,利用全连接对网络输出特征进行分类识别。与Baseline相比,STFP-Net引入了可忽略不计的额外参数和计算量。实验结果表明,与近些年主流方法相比,STFP-Net在主流数据库UCF101和HMDB51上的分类准确度具有明显提升。

关键词:行为识别;2D卷积网络;时空特征;特征金字塔;空洞卷积金字塔
文献标志码:A　**中图分类号:**TP391.4

Video Action Recognition Based on Spatio-Temporal Feature Pyramid Module

GONG Suming, CHEN Ying+
Key Laboratory of Advanced Process Control for Light Industry, Ministry of Education, Jiangnan University, Wuxi, Jiangsu 214122, China

Abstract: At present, the mainstream 2D convolution neural network method for video action recognition can't extract the relevant information between input frames, which makes it difficult for the network to obtain the spatio-temporal feature information between input frames and improve the recognition accuracy. To solve the existing problems, a universal spatio-temporal feature pyramid module (STFPM) is proposed. STFPM consists of feature pyramid and dilated convolution pyramid, which can be directly embedded into the existing 2D convolution network to form a new action recognition network named spatio-temporal feature pyramid net (STFP-Net). For multi-frame image input, STFP-Net first extracts the individual spatial feature information of each frame input and records it as the original feature. Then, the designed STFPM uses matrix operation to construct the feature pyramid of the original feature. Furthermore, the spatio-temporal features with temporal and spatial correlation are extracted by applying the dilated convolution pyramid to feature pyramid. Next, the original features and spatio-temporal features are fused by a weighted summation and transmitted to the deep network. Finally, the action in the video is classified by full connected layer. Compared with Baseline, STFP-Net introduces negligible additional parameters and computational complexity. Experimental results show that compared with mainstream methods in recent years,

基金项目:国家自然科学基金(61573168)。
This work was supported by the National Natural Science Foundation of China (61573168).
收稿日期:2020-12-31　**修回日期:**2021-02-25

STFP-Net has significant improvement in classification accuracy on the general datasets UCF101 and HMDB51.
Key words: action recognition; 2D convolution network; spatio-temporal features; feature pyramid; dilated convolution pyramid

在计算机视觉领域，对人类行为识别[1]的研究既能发展相关理论基础，又能扩大其工程应用范围。对于理论基础，行为识别领域融合了图像处理、计算机视觉、人工智能、人体运动学和生物科学等多个学科的知识，对人类行为识别的研究可以促进这些学科的共同进步。对于工程应用，视频中的人类行为识别系统有着丰富的应用领域和巨大的市场价值。其应用领域包括自动驾驶、人机交互、智能安防监控等。

早期的行为识别方法主要依赖较优异的人工设计特征，如密集轨迹特征[2]、视觉增强单词包法[3]等。得益于神经网络的发展，目前基于深度学习的行为识别方法已经领先于传统的手工设计特征的方法。Karpathy等[4]率先将神经网络运用于行为识别，其将单张RGB图作为网络的输入，这只考虑了视频的空间表观特征而忽略了时域上的运动信息。对此，Simonyan等[5]提出了双流网络。该方法使用基于RGB图片的空间流卷积神经网络（spatial stream ConvNet）和基于光流图的时间神经网络（temporal stream ConvNet）分别提取人类行为的静态特征和动态特征，最后将双流信息融合进行识别。Wang等[6]提出了时间片段网络（temporal segment network，TSN）结构来处理持续时间较长的视频，其将一个输入视频分成K段（segment），然后每个段中随机采样得到一个片段（snippet）。不同片段的类别得分采用段共识函数进行融合来产生段共识，最后对所有模型的预测融合产生最终的预测结果。为了解决背景信息干扰，周波等[7]结合目标检测使神经网络有侧重地学习人体的动作信息。刘天亮等[8]提出融合卷积网络（convolutional neural network，CNN）与长短期记忆网络（long short-term memory，LSTM）的行为识别方法来指导网络学习更有效的特征。借鉴2D卷积神经网络在静态图像的成功，Ji等[9]将2D卷积拓展为3D卷积，从而提出了3D-CNN方法来提取视频中的运动信息。Qiu等[10]将3D卷积解耦为2D空间卷积和1D时间卷积，在一定程度上减少了网络参数，缓解了网络难以优化的问题。Zhou等[11]提出了结合3D和2D的想法，其核心思想是在空间2D卷积网络中，加入3D卷积核（$T\times1\times1$）来获取视频序列中多帧之间的相关信息，以此来补充时间维度上的特征。

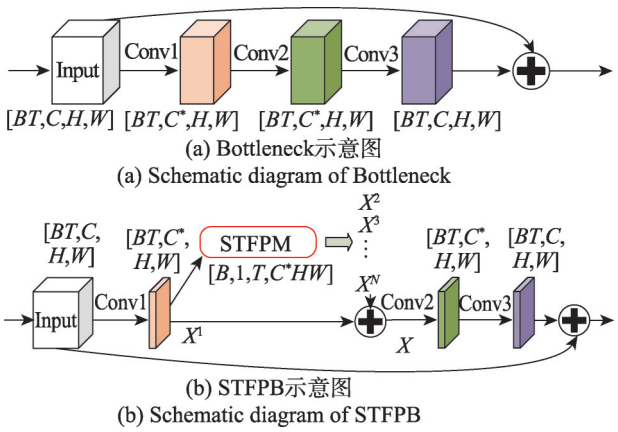
上述方法中，普通的2D卷积网络无法学习输入帧间时空特征信息，从而导致整体结果不佳；3D卷积方法虽然能同时提取表观信息和时空特征信息，但网络参数过多致使网络难以优化。基于此，本文提出全新的时空特征金字塔模块，该模块对输入特征构建特征金字塔模型并使用空洞卷积[12]金字塔提取输入帧间时空特征信息，同时只引入较少的额外参数和计算量。该模块是一种即插即用模块，能嵌入主流2D卷积网络中提升识别精度。

1 时空特征金字塔网络

本章首先介绍残差网络ResNet50[13]的构成模块Bottleneck；接着介绍将特征金字塔模块引入Bottleneck后的网络构成模块；最后给出本文网络的整体架构。特征金字塔模块将在2.1节详细介绍。

1.1 ResNet50基础模块问题分析

ResNet50由4个网络层（Stage1~4）构成，每层构成模块是Bottleneck，其结构过程如图1（a）所示。该模块首先对输入特征图使用1×1卷积（Conv1）进行卷积操作，主要目的是为了减少参数的数量，从而减少计算量，且在降维之后可以更加有效、直观地进行数据的训练和特征提取。接着使用感受野较大的3×3卷积（Conv2）来提取更细化特征。最后使用1×1卷积（Conv3）进行升维操作，使输出能与原始输入维度相匹配，从而进行特征相加。



(a) Bottleneck示意图
(a) Schematic diagram of Bottleneck

(b) STFPB示意图
(b) Schematic diagram of STFPB

图1 ResNet50构成模块及改进模块

Fig.1 Composition module of ResNet50 and corresponding improvement module

从图中可以看到，模块输入尺寸为 $[BT,C,H,W]$，B 表示批大小（batch_size），T 表示输入帧数，C 表示通道数，H、W 则是特征图尺寸大小。尽管网络输入帧数为 T，但维度 T 一直与维度 B 在一起，因此网络依旧是对每帧输入进行单独操作，并未提取帧间信息。

为了解决Bottleneck存在的问题，本文提出了时空特征金字塔模块并将其插入Bottleneck构建STFPB（spatio-temporal feature pyramid block），该模块流程如图1(b)所示。

输入特征 $[BT,C,H,W]$ 首先经过1次卷积操作变成 $[BT,C^*,H,W]$，记为 X^1；然后将 X^1 送入STFPM（spatio-temporal feature pyramid module）提取时空特征信息，在此部分特征尺寸将变为 $[B,1,T,C^*HW]$，得到时空特征后将特征尺寸转化为原始尺寸；然后将 X^1、X^2、X^3 等加权融合，其中 X^1 权值固定为1，X^2、X^3 等权值由网络学习得到；最后将融合特征 X 送入后续的两个卷积层并与原始输入特征相加。

1.2 网络整体架构

以STFPB为基础，本文构建了全新的行为识别网络STFP-Net。网络整体结构如图2所示。图中，黑色虚线框表示网络层（stage），紫红色立方体表示输入，黄色矩形表示卷积核为7×7的卷积操作，蓝色立方体表示ResNet50原本的Bottleneck（①~⑬），红色立方体表示STFPB（⑭~⑯），绿色框表示全连接层（Fc），橙色椭圆表示交叉熵损失函数，紫色圆圈则是每类结果的预测得分。输入首先经过7×7卷积操作进行特征图尺寸缩减，然后将输出送入Bottleneck与STFPB进行特征提取，最后的高层特征经过全连接层拉平后得到每类结果的预测得分，完成行为识别任务。

2 时空特征金字塔模块

本章首先介绍特征金字塔，其作用是说明需要对哪些输入帧进行操作；接着介绍空洞卷积和由它构成的空洞卷积金字塔，其作用是提取时空特征信息；最后介绍时空特征的融合方式。

2.1 特征金字塔

给定一个输入 $F\in\mathbb{R}^{B\times C\times T\times H\times W}$，首先通过网络提取特征，然后如图3所示构建特征金字塔。图3中，蓝色立方体表示某一帧图像的全部特征信息，紫色虚线立方体表示该帧被跳过，不参与特征金字塔的构建。为了画图简便，省略了维度 B，同时 T_i 表示第 i 帧的全部特征信息。

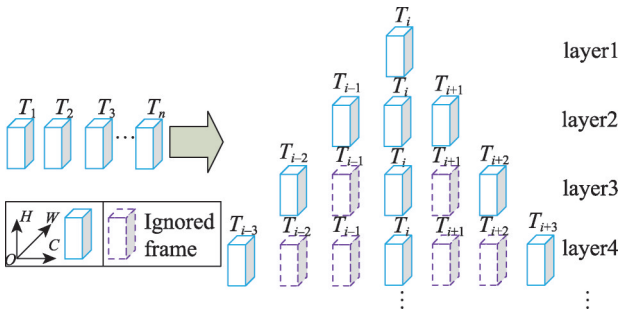


图3 特征金字塔示意图

Fig.3 Schematic diagram of feature pyramid

以 T_i 为例，在特征金字塔中，T_i 作为金字塔的第一层；然后 (T_{i-1},T_i,T_{i+1}) 三帧图像特征信息作为金字塔第二层；接着将采样步长设为2，则金字塔第三层便是 (T_{i-2},T_i,T_{i+2})，此时跳过了 (T_{i-1},T_{i+1})。以此类推，便能构建多层特征金字塔。值得注意的是，金字塔第一层只包含 T_i，后续的每一层都包含3帧图像特征信息。

此外，和其他特征金字塔方法不同的是，本文的特征金字塔并不减小特征图的尺寸，这避免了原始

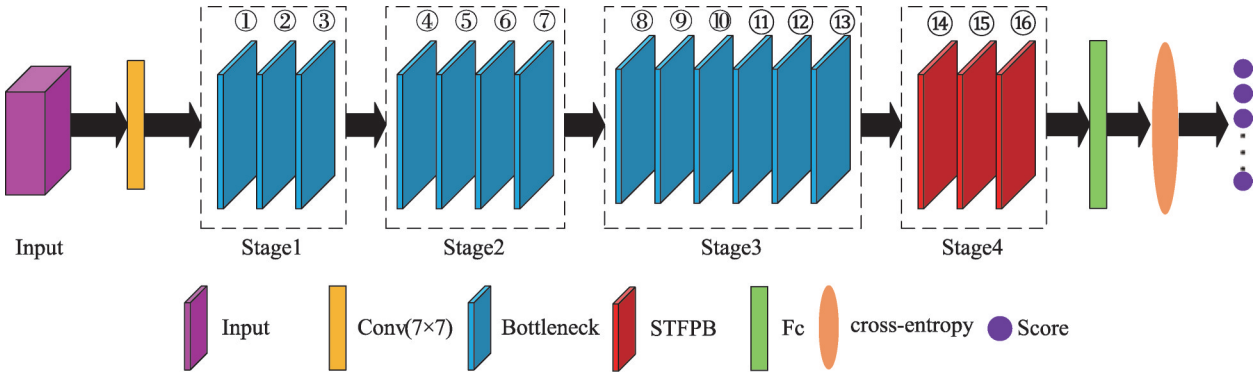


图2 网络整体架构

Fig.2 Overall structure of network

信息的丢失。与此同时，本文也不引入额外的损失函数来指导网络学习，这从网络的训练和复杂性角度来说是有益的。

2.2 空洞卷积金字塔和时空特征提取

与普通卷积相比，空洞卷积多了一个超参数dilation factor。图4给出了空洞卷积的示意图。当dilation factor等于1时，此时的卷积核便是普通的卷积核，随着dilation factor的增大，卷积核的尺寸也在变大。图中红点表示卷积核需要学习的参数，其余的空白部分用0进行填充。

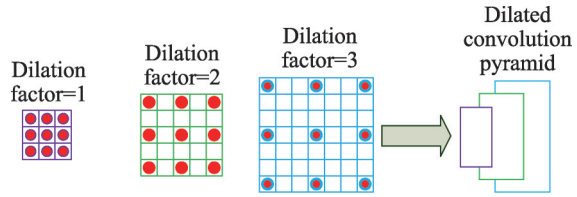


图4 空洞卷积和卷积金字塔

Fig.4 Dilated convolution and convolution pyramid

空洞卷积的主要优势是在特征图不做池化操作损失信息的前提下，加大了感受野，让每个卷积输出都包含较大范围的信息。此外，空洞卷积的另一个优势便是0值填充带来的“跳跃”特性。很多文章都没有考虑过此特性，本文将不同dilation factor的空洞卷积组合在一起，构建了空洞卷积金字塔，并将它应用到特征金字塔中提取帧间时空特征信息。

完整的时空特征提取过程如图5①②所示。以 T_i 为例，将它当作特征金字塔的第一层，并记为 X_i^1；接着，将dilation factor等于1的空洞卷积作用在金字塔第二层 (T_{i-1},T_i,T_{i+1})，将提取到的时空特征记为 X_i^2；然后，对第三层金字塔（T_{i-2},T_i,T_{i+2}）使用dilation factor等于2的空洞卷积提取时空特征并记为 X_i^3；总之，第 N 层的金字塔特征 X_i^N 可以通过在 F^* 上使用dilation factor等于 $N-1$ 的空洞卷积来获得。通过上述操作便能得到时空特征 $(X_i^1,X_i^2,\cdots,X_i^N)$，接着将时空特征送入后续特征融合操作。

2.3 时空特征的融合策略

对得到的时空特征，本文考虑了两种特征融合策略：特征级联和加权相加。

假设有两路通道数相同的输入 $(Y_1,Y_2,\cdots,Y_c)$ 和 $(Z_1,Z_2,\cdots,Z_c)$，特征级联和加权相加分别为：

$$F_{\text{con}}=\sum_{i=1}^{c}Y_i\times K_i+\sum_{i=1}^{c}Z_i\times K_{i+c} \tag{1}$$

$$F_{\text{add}}=\sum_{i=1}^{c}(\alpha Y_i+\beta Z_i)\times K_i \tag{2}$$

式(1)、式(2)中，Y 表示原始特征，Z 表示提取后的特征，c 表示通道数，K 表示卷积操作，α 和 β 表示权重系数。

可以这么理解上述公式，特征级联是通道数的合并，也就是说描述图像本身的特征数（通道数）增加了，而每一特征下的信息没有增加。加权相加则是每一维下的特征信息量在增加，特征数（通道数）没有改变。特征级联操作中每个通道对应着相应的卷积核，而加权相加操作则将对应的特征图相加，再进行下一步卷积操作，相当于加了一个先验：对应通道的特征图语义类似，从而对应的特征图共享一个卷积核。本文的核心思想是对空域特征进行时域信息的增加，这更符合特征加权相加的思想，因此，本文选用加权相加的方式进行特征融合，后续对比实验也佐证了这一想法。

本文中，第一层金字塔特征的权重固定为1，其余各层金字塔特征的权重系数由网络学习得到。

2.4 时空特征金字塔模块具体流程

完整的时空特征金字塔流程如图5所示。给定一个输入 $F\in\mathbb{R}^{B\times C\times T\times H\times W}$，首先通过矩阵维度变换操作，将 F 变为 $F^*\in\mathbb{R}^{B\times 1\times T\times CHW}$（图5①）。图5①中，矩形的每一行表示某一帧的全部特征。然后对 F^* 使用

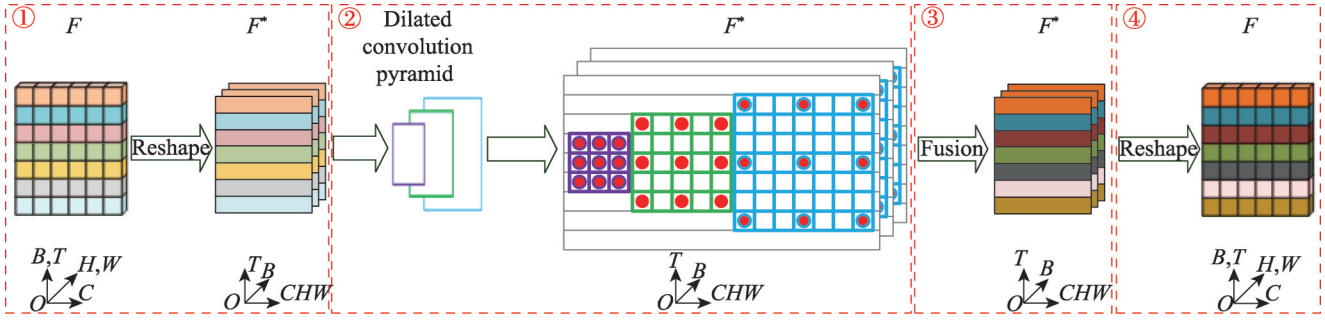


图5 时空特征金字塔详细过程

Fig.5 Detailed process of spatio-temporal feature pyramid

空洞卷积金字塔进行时空特征提取（图5②），接着对得到的时空特征使用加权相加策略进行特征融合操作（图5③），此时融合后的特征的尺寸和原始输入不相符，因此需要再次使用矩阵维度转换操作将特征维度进行转换（图5④）。相比原始输入，经过时空特征金字塔模块后的特征具有时空特征信息，因此更有利于视频行为识别任务。

3 实验验证和分析

3.1 数据集介绍

本文在最常见的行为识别数据集UCF101[14]和HMDB51[15]上对本文网络结构进行评估实验，以便将其性能与目前主流的方法进行比较。

UCF101数据集是从YouTube收集的具有101个动作类别的逼真动作视频的动作识别数据集。101个动作类别中的视频分为25组，每组可包含4~7个动作视频。来自同一组的视频可能共享一些共同的功能，例如类似的背景、类似的观点等。

HMDB51数据集内容主要来自电影，一小部分来自公共数据库，如YouTube视频。该数据集包含6 849个剪辑，分为51个动作类别，每个动作类别至少包含101个剪辑。

3.2 实验设置

3.2.1 Baseline设置

本文Baseline采用ResNet50作为主干网络。针对每个视频输入，首先将其分为8个片段，然后在每个片段随机采样1帧，共计8帧作为输入。网络对每帧作出预测，然后将8个预测值取平均作为最终预测值。

3.2.2 训练设置

本文实验中，卷积神经网络基于PyTorch平台设计实现。网络采用ResNet50作为主干网络，训练采用小批量随机梯度下降法，动量为0.9，权值在第15、35、55个epoch时衰减一次，衰减率为0.1，总训练数设置为70。初始学习率设为0.001。Dropout设置为0.8。实验采用两张TITAN 1080Ti GPU进行，batch_size设置为24。

3.3 实验结果与分析

3.3.1 特征融合策略

本文主要考虑两种特征融合策略：特征级联和加权融合。表1给出了两种融合策略在UCF101上的结果。从结果来看，加权融合比特征级联高了3.331个百分点，故本文后续实验均采用加权融合方式。在2.3节中已经说明，金字塔第一层特征权值固定为1，后续层的特征权值由网络自主学习得到。

表1 特征融合策略的影响

Table 1 Influence of feature fusion strategy

融合策略	UCF101正确率/%
特征级联	86.994
加权融合	90.325

3.3.2 金字塔层数的影响

从2.1节和2.2节描述可以知道，特征金字塔的层数和空洞卷积金字塔的层数是相同的。理论上，当输入帧是无限数量时，可以构建无限多层金字塔。因此采用ResNet50作为主干网络，然后用实验来说明金字塔层数最合理的值。实验结果如表2所示。从结果来看，金字塔层数为3时结果最高。主要原因如下：2层金字塔只包含两个相邻帧的信息，这限制了它提取时间信息的能力；当层数大于等于4时，空洞卷积本身的“网格”效应造成严重的信息不连续，影响最终的识别结果。因此，在后续实验中，金字塔等级的数量固定为3。

表2 金字塔层数的影响

Table 2 Influence of pyramid layers

层数	UCF101正确率/%
2	88.686
3	90.325
4	89.638

3.3.3 金字塔模块嵌入位置的分析

金字塔模块可以直接嵌入到现有网络中使用，选择合适的嵌入位置对网络来说至关重要。对于深度神经网络，低层网络会提取一些边缘特征，然后高层网络进行形状或目标的认知，更高的语义层会分析一些运动和行为。

基于此共识，本文首先在ResNet50的Stage4中添加金字塔模块，然后逐步往低层网络添加金字塔模块。表3给出了不同位置嵌入金字塔后的结果。从结果来看，在语义层（Stage4）嵌入金字塔模块后表现最好。当Stage3和Stage4同时嵌入金字塔模块后，识别结果下降了0.106个百分点，这是因为行为识别任务更依赖语义信息。因此，本文最终只在Stage4中嵌入金字塔模块。

表3 嵌入位置的影响

Table 3 Influence of embedding position

嵌入位置	UCF101正确率/%
Stage{3, 4}	90.219
Stage4	90.325

3.3.4 网络参数和计算量分析

通过前3个小实验，本文网络最终配置为3层金

字塔,Stage4添加和加权融合,并记为STFP-Net。一个优秀的嵌入性模块不仅能给网络带来结果正确率的提升,同时引入的额外计算量也应该很少。基于此,对本文最终网络的模型大小与计算量进行定量分析,如表4所示。采用每秒浮点运算次数(FLOPs)作为计算量的评价指标,该指标值越大则意味着网络需要更多的计算资源。

表4 模型参数和计算量

Table 4 Model parameters and calculation amount

网络	模型大小/MB	FLOPs/10^9
Baseline	25.557 032	32.892 117
STFP-Net(2层)	25.557 059	32.902 955
STFP-Net(3层)	25.557 086	32.913 793
STFP-Net(4层)	25.557 113	32.924 631

从表4结果来看,相比于Baseline,3层金字塔只增加了54 Byte的模型参数,计算量只增加了0.06%,约为2×10^7次浮点运算数量。综上所述,本文的金字塔模块是高效的嵌入性模块。

3.3.5 与主流方法比较

在本小节中,将通过具体实验进一步展示所提出的STFP-Net与最先进的动作识别方法的比较结果。关于UCF101和HMDB51的相关结果详见表5。

表5 与主流方法正确率的比较

Table 5 Comparison of accuracy with mainstream methods 单位:%

方法	UCF101	HMDB51	方法	UCF101	HMDB51
C3D+IDT[9]	90.4	—	I3D[21]	95.7	74.3
R(2+1)D[16]	95.0	72.7	StNet[22]	93.5	—
LTC[17]	91.7	64.8	Hidden-ts[23]	93.2	66.8
Ts+LSTM[18]	88.6	—	STH[24]	96.0	74.8
TLE[19]	95.4	71.1	ISTPA-Net[25]	95.5	70.7
MiCT[11]	94.7	70.5	Baseline	94.2	69.4
TSM[20]	94.5	70.7	Ours(STFP-Net)	**96.4**	**75.5**
P3D[10]	93.7	—			

在当下主流方法中,有的采用了先进的时空融合方法来获得高效的网络特征,如TLE[19];有的则利用CNN和LSTM网络的结合体来获得输入帧之间的序列信息以此来获得比单纯RGB表观信息更丰富的时空信息;I3D[21]直接将最先进的2D CNN架构膨胀成3D CNN网络,以利用训练好的2D模型;为了减少参数量,P3D[10]通过将3D卷积分解为沿空间维度的2D卷积和沿时间维度的1D卷积来建模时空信息,从而学习非常深的时空特征;MiCT[11]则提出混合2D/3D卷积模块,利用2D卷积提取RGB图像的表观信息,利用3D卷积提取序列间的相关信息。ISTPA-Net[25]通过对不同网络层的特征进行下采样从而构建注意力金字塔,旨在突出特征图中某些重要的特征,同时引入两个额外的损失函数来约束网络学习。

从表5结果来看,在UCF101数据集上,本文的STFP-Net以96.4%的正确率排在所有方法中第一位;同样的,在HMDB51数据集上,本文的STFP-Net以75.5%的正确率排在第一名。

综上所述,由本文提出的特征金字塔模块所构建的STFP-Net确实具有明显的效果提升。

4 结束语

本文提出了时空特征金字塔模块下的人体行为识别方法。通过分析现有基础网络的局限性,提出了时空特征金字塔模块。为了验证模块的有效性,分别从特征融合方式、金字塔层数、嵌入位置、额外计算量等方面进行实验验证。最后在通用数据集上与其他主流方法进行比较,实验结果再次证明了时空特征金字塔模块的高效性。

参考文献:

[1] 朱红蕾, 朱昶胜, 徐志刚. 人体行为识别数据集研究进展[J]. 自动化学报, 2018, 44(6): 978-1004.
ZHU H L, ZHU C S, XU Z G. Research progress of human action recognition datasets[J]. Acta Automatica Sinica, 2018, 44(6): 978-1004.

[2] IKIZLER-CINBIS N, SCLAROFF S. Object, scene and actions: combining multiple features for human action recognition[C]//LNCS 6311: Proceedings of the 11th European Conference on Computer Vision, Heraklion, Sep 5-11, 2010. Berlin, Heidelberg: Springer, 2010: 494-507.

[3] 张良, 鲁梦梦, 姜华. 局部分布信息增强的视觉单词描述与动作识别[J]. 电子与信息学报, 2016, 38(3): 549-556.
ZHANG L, LU M M, JIANG H. An improved scheme of visual words description and action recognition using local enhanced distribution information[J]. Journal of Electronics & Information Technology, 2016, 38(3): 549-556.

[4] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Jun 23-28, 2014. Washington: IEEE Computer Society, 2014: 1725-1732.

[5] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Proceedings of the Annual Conference on Neural Information Processing Systems 2014, Montreal, Dec 8-11, 2014. Red Hook: Curran Associates, 2014: 568-576.

[6] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//LNCS 9912: Proceedings of the 14th European

Conference on Computer Vision, Amsterdam, Oct 8-16, 2016. Cham: Springer, 2016: 20-36.

[7] 周波, 李俊峰. 结合目标检测的人体行为识别[J]. 自动化学报, 2019, 42(5): 56-67.
ZHOU B, LI J F. Human action recognition combined with object detection[J]. Acta Automatica Sinica, 2019, 42(5): 56-67.

[8] 刘天亮, 谯庆伟, 万俊伟, 等. 融合空间-时间双网络流和视觉注意的人体行为识别[J]. 电子与信息学报, 2018, 40(10): 2395-2401.
LIU T L, QIAO Q W, WAN J W, et al. Human action recognition based on spatial-temporal double network flow and visual attention[J]. Journal of Electronics and Information Technology, 2018, 40(10): 2395-2401.

[9] JI S W, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[10] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017. Washington: IEEE Computer Society, 2017: 5533-5541.

[11] ZHOU Y, SUN X, ZHA Z J, et al. MICT: mixed 3D/2D convolutional tube for human action recognition[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 449-458.

[12] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[C]//Proceedings of the 4th International Conference on Learning Representations, San Juan, May 2-4, 2016: 1-13.

[13] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 770-778.

[14] SOOMRO K, ROSHAN ZAMIR A, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. arXiv:1212.0402, 2012.

[15] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Nov 6-13, 2011. Washington: IEEE Computer Society, 2011: 2556-2563.

[16] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018. Washington: IEEE Computer Society, 2018: 6450-6459.

[17] 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.

[18] NG Y H, HAUSKNECHT M J, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification[C]//Proceedings of the 2015 IEEE Confer-ence on Computer Vision and Pattern Recognition, Boston, Jun 7-12, 2015. Washington: IEEE Computer Society, 2015: 4694-4702.

[19] DIBA A, SHARMA V, VAN GOOL L. Deep temporal linear encoding networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 1541-1550.

[20] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 7082-7092.

[21] 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, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 4724-4733.

[22] HE D L, ZHOU Z C, GAN C, et al. StNet: local and global spatial-temporal modeling for action recognition[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence, the 31st Innovative Applications of Artificial Intelligence Conference, the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, Jan 27-Feb 1, 2019. Menlo Park: AAAI, 2019: 8401-8408.

[23] ZHU Y, LAN Z Z, NEWSAM S D, et al. Hidden two-stream convolutional networks for action recognition[C]//LNCS 11363: Proceedings of the 14th Asian Conference on Computer Vision, Perth, Dec 2-6, 2018. Cham: Springer, 2018: 363-378.

[24] LI X, WANG J, MA L, et al. STH: spatio-temporal hybrid convolution for efficient action recognition[J]. arXiv:2003.08042, 2020.

[25] DU Y, YUAN C F, LI B, et al. Interaction-aware spatio-temporal pyramid attention networks for action classification[C]//LNCS 11220: Proceedings of the 15th European Conference on Computer Vision, Munich, Sep 8-14, 2018. Cham: Springer, 2018: 388-404.

龚苏明(1995—),男,江苏镇江人,硕士研究生,主要研究方向为模式识别、行为识别。
GONG Suming, born in 1995, M.S. candidate. His research interests include pattern recognition and action recognition.

陈莹(1976—),女,浙江丽水人,博士,教授,CCF高级会员,主要研究方向为模式识别、信息融合。
CHEN Ying, born in 1976, Ph.D., professor, senior member of CCF. Her research interests include pattern recognition and information fusion.