计算机科学与探索
1673-9418/2000/00(00)-00000-00
Journal of Frontiers of Computer Science and Technology doi: 10.3778/j.issn.1673-9418.0000000

时空特征金字塔模块下的视频行为识别

龚苏明[1], 陈莹[1+]

1. 江南大学 轻工过程先进控制教育部重点实验室，江苏省 无锡市 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，导致网络无法获得时空特征信息进而难以提升精度．针对目前存在的问题，提出了通用的时空特征金字塔模块(Spatio-temporal Feature Pyramid Module, 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卷积网络中构成新的行为识别网络(STFP-Net)．针对多帧图像输入，STFP-Net首先提取每帧输入的单独特征信息，记为原始特征；然后，所设计的STFPM利用矩阵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深层网络；最后，利用全连接对视频中行为进行分类．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：TP***

Video Action Recognition based on Spatio-temporal Feature Pyramid Module

GONG Suming[1], CHEN Ying[1+]

1. 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 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 and improve the 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, and can be directly embedded into the existing 2D convolution network to form a new action recognition network (STFP-Net). For multi-frame image input, STFP-Net firstly extracts the individual 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. Secondly, the spatio-temporal features with temporal and spatial correlation are extracted by using the dilated convolution pyramid. Then, 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

基金项目：国家自然科学基金项目（编号: 61573168）。

This work was supported by the National Natural Science Foundation of China under Grant (No: 61573168).

computational complexity. The experimental results show that compared with the mainstream methods in recent years, STFP-Net has a 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

1 引言

在计算机视觉领域, 对人类行为识别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[1]、视觉增强单词包法[2]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[3]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[4]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[5]提出了 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, Zhou 等[6]结合目标检测使神经网络有侧重地学习人体的动作信息. Liu 等[7]提出融合卷积网络与 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[8]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[9]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[10]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($T \times 1 \times 1$)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.

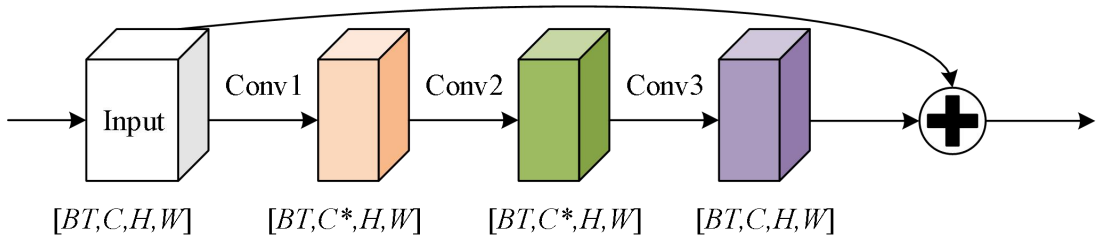
上述方法中, 普通的 2D 卷积网络无法学习输入帧间时空特征信息, 从而导致整体结果不佳; 3D 卷积方法虽然能同时提取表观信息和时空特征信息, 但网络参数过多致使网络难以优化. 基于此, 本文提出全新的时空特征金字塔模块, 该模块对输入特征构建特征金字塔模型并使用空洞卷积[11]金字塔提取输入帧间时空特征信息, 同时只引入较少的额外参数和计算量. 该模块是一种即插即用模块, 能嵌入主流 2D 卷积网络中提升识别精度.

2 时空特征金字塔网络

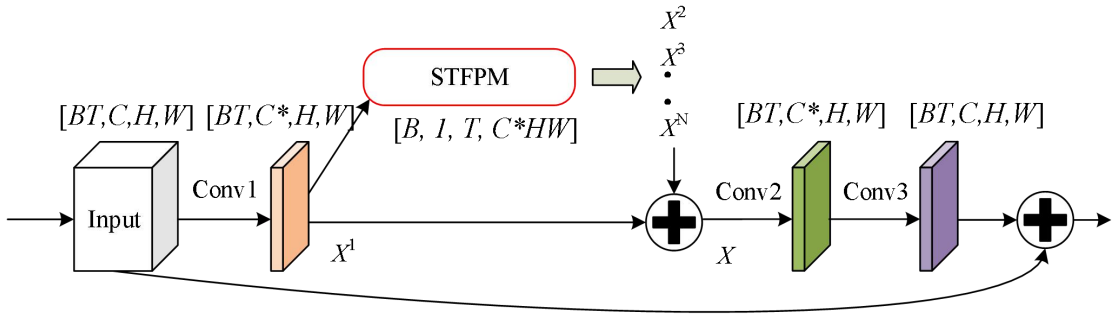
本节首先介绍 ResNet50[12]的构成模块—Bottleneck; 接着介绍将特征金字塔模块引入 Bottleneck 后的网络构成模块; 最后给出本文网络的整体架构. 特征金字塔模块将在第 3 节详细介绍.

2.1 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.



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



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

Fig.1 Composition module of ResNet50 and corresponding improvement module

图 1 ResNet50 构成模块及改进模块

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.

在图中可以看到，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 Spatio-temporal Feature Pyramid Block(STFPB)，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 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 2 个卷积层并与原始输入特征相加.

2.2 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 模块，红色立方体表示 STFPB，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.

3 时空特征金字塔模块

本小节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

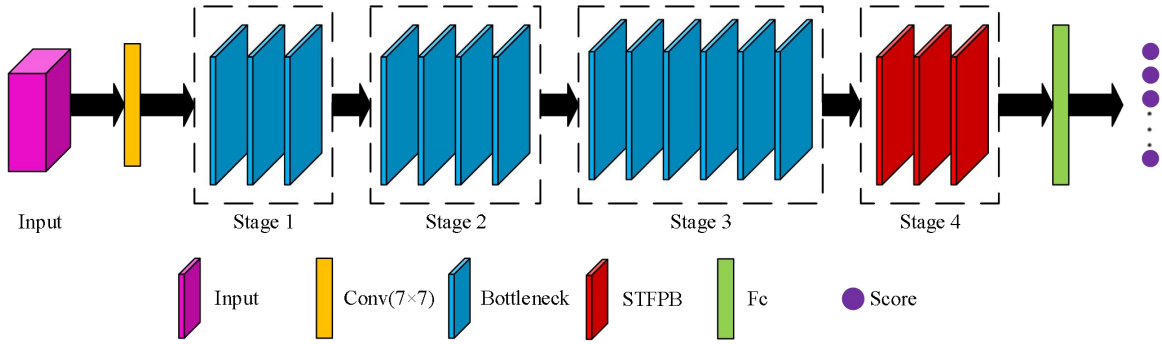


Fig.2 Overall structure of network

图 2 网络整体架构

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.

3.1 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.

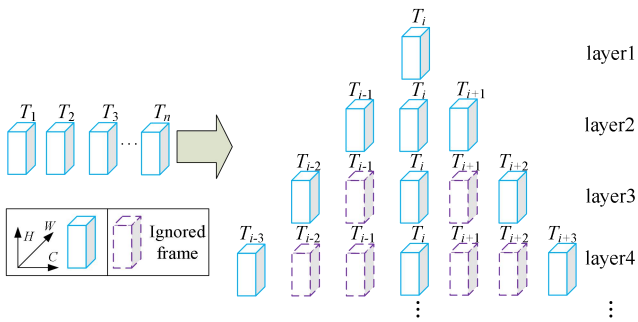


Fig.3 Schematic diagram of feature pyramid

图 3 特征金字塔示意图

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.

3.2 空洞卷积金字塔和时空特征提取

与普通卷积，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.

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.

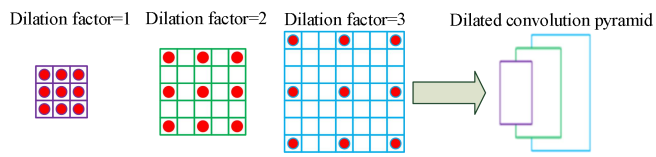


Fig.4 Dilated convolution and convolution pyramid

图 4 空洞卷积和卷积金字塔

完整的时空特征提取过程如图 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 5 ①②). 图 5②中，矩形的每一行表示某一帧的全部特征.

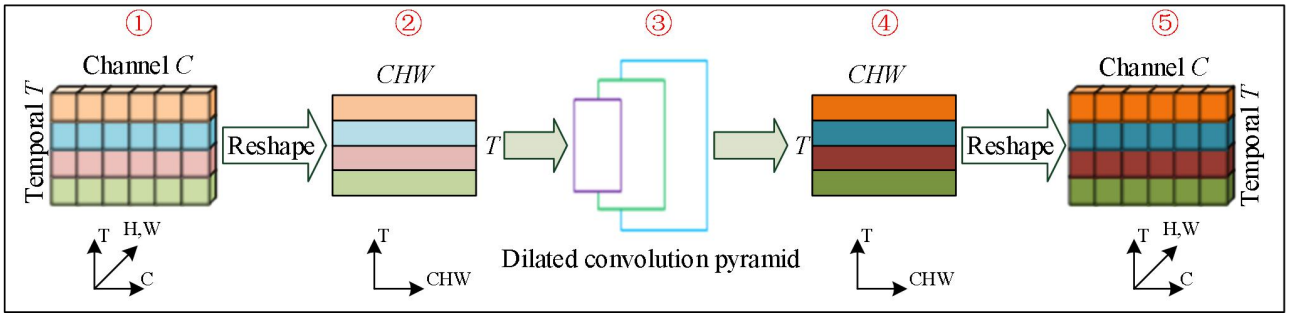


Fig.5 Extraction process of spatio-temporal features

图 5 时空特征提取过程

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．将得到的时空特征(X_i^1, X_i^2, … , X_i^N)进行融合，此时的 T_i 便包含不同帧之间的时空特征．对所有输入帧进行上述操作，原始的单帧特征就变成了包含帧间时空特征信息的全新特征．最后通过矩阵操作将新特征的维度转换为原始输入的维度从而进行后续特征学习．图 5②③④⑤演示了上述步骤．

3.2 时空特征的融合策略

对得到的时空特征，本文考虑了 2 种特征融合策略：特征级联和加权相加．

假设已经得到 N 层金字塔时空特征(X_i^1, X_i^2, … , X_i^N)，每个特征的尺寸为 $\mathbb{R}^{BT\times C\times H\times W}$．通过特征级联操作，得到一个新特征 $S\in\mathbb{R}^{BT\times CN\times H\times W}$，然后对 S 使用卷积核为 1×1 的 2D 卷积进行特征融合和维度减小，并将它送入后续网络．

从特征金字塔的构建过程来看，提取的时空信息可以看作是对当前帧特征的增强，因此本文使用带权重的逐元素加法来代替普通的逐元素加法．本文中，X_i^1 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．

4 实验验证和分析

4.1 数据集介绍

本文在最常见的行为识别数据集 UCF101[13] 和 HMDB51[14] 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 6849 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．

4.2 实验设置

4.2.1 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．

4.2.2 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 2 张 TITAN 1080TI GPU 进行, batchsize 设置为 24.

4.3 实验结果与分析

4.3.1 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 3.4 个百分点，故本文后续实验均采用加权融合方式. 在 3.2 小节中已经说明，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.

Table 1 Influence of feature fusion strategy

表 1 特征融合策略的影响

融合策略	UCF101(%)
特征级联	86.994
加权融合	90.325

4.3.2 金字塔层数的影响

从第三节描述可以知道，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. 所以我们采用 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.

Table 2 Influence of pyramid layers

表 2 金字塔层数的影响

层数	UCF101(%)
2	88.686
3	90.325
4	89.638

4.3.3 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 0.11 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.

Table 3 Influence of embedding position

表 3 嵌入位置的影响

嵌入位置	UCF101(%)
stage{3, 4}	90.219
Stage4	90.325

4.3.4 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. 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.

从表 4 结果来看，相比于 Baseline, 3 层金字塔只增加了 54B 的模型参数，计算量只增加了 0.06%的计算量，约为 0.02GB. 综上所述，本文的金字塔模块是高效的嵌入性模块.

Table 4 Model parameters and calculation amount

表 4 模型参数和计算量

网络	模型大小(MB)	FLOPs(GB)
Baseline	25.557032	32.892117
STFP-Net(2 层)	25.557059	32.902955
STFP-Net(3 层)	25.557086	32.913793
STFP-Net(4 层)	25.557113	32.924631

4.3.5 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.

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[15]；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[16]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[9]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[10]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.

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.

Table 5 Comparison with mainstream methods

表 5 与主流方法的比较

方法	UCF101(%)	HMDB51(%)
C3D+IDT[8]	90.4	-
R(2+1)D[17]	95.0	72.7
LTC[18]	91.7	64.8
Ts+LSTM[19]	88.6	-
TLE[15]	95.4	71.1
MiCT[10]	94.7	70.5
TSM[20]	94.5	70.7
P3D[9]	93.7	-
I3D[16]	95.7	74.3
StNet[21]	93.5	-
Hidden-ts[22]	93.2	66.8
STH[23]	96.0	74.8
Baseline	94.2	69.4
Ours(STFP-Net)	**96.4**	**75.5**

5 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] IKIZLER C N, SCLAROFF S. Object, sence and actions: Combining multiple features for human action recognition[C]// European Conference on Computer Vision, Heraklion, Crete, Greece, Sep 5-11, 2010. Springer, 2010, 6311: 494-507.

[2] 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.
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.

[3] KARPATHY A, TODERICI G, Shetty S, et al. Large-scale video classi-fication with convolutional neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colombus, USA, Jun 23-28, 2014. IEEE, 2014: 1725-1732.

[4] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Proceedings of Advances in Neural Information Processing Systems, Montreal, Canada, Dec 8-11, 2014. IEEE, 2014: 568-576.

[5] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recogni-tion[C]// Proceedings of European Conference on Computer Vision, Amsterdam, Holland. Oct 8-16, 2016. Springer, 2016: 20-36.

[6] Zhang B B, Ge S Y, Wang Q L, et al. Multi-order Information Fusion Method for Human Action Recogni-tion[J/OL]. Acta Automatica Sinica.
张冰冰，葛疏雨，王旗龙，李培华. 基于多阶信息融合的行为识别方法研究[J/OL]. 自动化学报.

[7] 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.
刘天亮，谯庆伟，万俊伟，戴修斌，罗杰波. 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.

[8] 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, Santiago, Chile, Dec 7-13, 2015. IEEE, 2015: 4489-4497.

[9] 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, Venice, Italy, Oct 22-29, 2017. IEEE, 2017: 5533-5541.

[10] 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, Salt Lake City, Utah, Jun 18-22, 2018. IEEE, 2018: 449-458.

[11] Yu F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C]// International Conference on Learning Representations, San Juan, Puerto Rico, May 2-4, 2016.

[12] 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, Las Vegas, NV, USA, Jun 27-30. IEEE. 2016: 770–778.

[13] Soomro K, Zamir, A R, et al. Ucf101: A dataset of 101 human actions classes from videos in the wild[J/OL]. arXiv preprint arXiv:1212.0402 (2012).

[14] Kuehne H, Jhuang H, Garrote E, et al. Hmdb: a large video database for human motion recognition[C]// 2011 International Conference on Computer Vision, Barcelona, Spain, Nov 6-13, 2011. IEEE,pp.2556–2563.

[15] Diba A, Sharma V, Van Gool L. Deep temporal linear encoding networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. IEEE, 2017: 2329-2338.

[16] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, Jul 21-26, 2017. IEEE, 2017: 6299-6308.

[17] 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, Salt Lake City, Utah, Jun 18-22, 2018. IEEE, 2018: 6450-6459.

[18] Varol G , Laptev I , Schmid C . Long-term Temporal Convolutions for Action Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018:1-1.

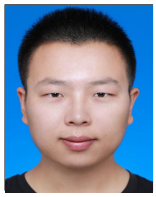
[19] Ng Y H , Hausknecht M , Vijayanarasimhan S , et al. Beyond short snippets: Deep networks for video classification[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, Jun 8-12, 2015. IEEE, 2015: 4694-4702.

[20] Lin J, Gan C, Han S. Tsm: Temporal shift module for efficient video understanding[C]// Proceedings of the IEEE International Conference on Computer Vision, Seoul, south Korea, Oct 27-Nov 2, 2019. IEEE, 2019: 7083-7093.

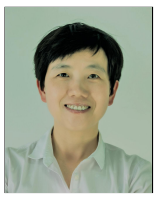
[21] He D, Zhou Z, Gan C, et al. Stnet: Local and global spatial-temporal modeling for action recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence, Hawaii, Jan 27-Feb 1, 2019. IEEE, 2019, 33: 8401-8408.

[22] Zhu Y, Lan Z, Newsam S, et al. Hidden two-stream convolutional networks for action recognition[C]// Asian Conference on Computer Vision, Perth, Australia, Dec 2-6. Springer, 2018: 363-378.

[23] Li X, Wang J, Ma L, et al. STH: Spatio-Temporal Hybrid Convolution for Efficient Action Recognition[J/OL]. arXiv preprint arXiv:2003.08042, 2020.



GONG Suming was born in 1995. He is an M.S. candidate at Jiangnan University. His main research direction is deep learning and pattern recognition.
龚苏明（1995—），男，江苏镇江人，江南大学硕士研究生，主要研究方向为深度学习、模式识别。



CHEN Ying was born in 1976. She received the Ph.D degree in engineering from Xi'an Jiaotong University in 2005. Now she is a professor and Ph.D. supervisor a at Jiangnan University. Her main research fields are pattern recognition and information fusion. So far, nearly 100 papers have been published in domestic and foreign journals and conferences as the first author or responsible author. He has presided over a number of provincial and ministerial level projects including the general program of National Natural Science Foundation of China, the youth fund project and the natural science foundation of Jiangsu Province.
陈莹（1976—）女， 浙江丽水人，2005 年西安交通大学获工学博士学位，江南大学教授，博士生导师。主要研究领域为模式识别，信息融合。目前为止，以第一作者或责任作者在国内外期刊及会议上公开发表论文近 100 篇，主持包括国家自然科学基金面上项目、青年基金项目以及江苏省自然科学基金在内的多项省部级以上项目，CCF 会员。