基于时间与通道混洗模块的视频行为识别

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　无锡　214122)
(chenying@jiangnan.edu.cn)

摘　要: 为了实现纯 2D 卷积神经网络提取视频帧之间的相关信息, 提出时间与通道混洗模块下的行为识别方法. 针对多帧图像输入, 主干网络首先提取每帧的单独信息, 记为原始信息; 然后, 所设计的时间与通道混洗模块利用矩阵操作将独立的输入特征图转换为具有时空关联性的全新特征图并提取融合信息, 记为时空信息; 接着, 将原始信息与时空信息进行相加并传递给深层网络; 最后, 利用全连接层对视频中行为进行分类. 实验结果表明, 与近些年主流方法相比, 该文方法在分类准确度上具有明显提升.

关键词: 行为识别; 2D 卷积神经网络; 时间与通道混洗

Video Action Recognition Based on Time and Channel Shuffling Module

Gong Suming, Chen Ying*

(*Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University Wuxi*　214122)

Abstract: In order to extract the relevant information between video frames by pure 2D convolutional neural network, an action recognition method based on time and channel shuffling module is proposed. For multi-frame image input, the backbone network first extracts the individual information of each frame and records it as the original information; Then, the designed time and channel shuffling module converts the independent input feature map into a new feature map with spatio-temporal correlation by matrix operation and extracts fusion information, which is recorded as spatio-temporal information; After that, the original information and spatio-temporal information are added and transmitted to the deep network; Finally, the action in the video is classified by full connected layer. Experimental results show that, compared with the mainstream methods in recent years, the classification accuracy of this method is obviously improved.

Key words: Action recognition; 2D convolution neural network; Time and channel shuffling

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

基金项目: **国家自然科学基金**(61573168). **龚苏明**(1995—), 男, 硕士研究生, 主要研究方向为行为识别; **陈莹**(1978—), 女, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为图像处理、计算机视觉.

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. 为了充分建模空间和时间特征分布, Zhang 等[6]提出一种基于二阶聚合的视频多阶信息融合方法. Shi 等[7]提出一种融合特征传播和时域分割网络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. 3D 卷积网络存在的主要问题是随着网络的加深, 参数量太过庞大, 因此难以对它进行优化. Qiu 等[9]将 3D 卷积解耦为 2D 空间卷积和 1D 时间卷积, 这样做一方面可以降低参数量, 另一方面可以使用现有的经典网络. 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 × 1 × 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. Tran 等[11]将 3D 卷积核(T × D × D)分解为“1 × D × D”和“T × 1 × 1”两种卷积核, 尽管不改变参数的数目, 但由于每个块中 2D 和 1D 卷积之间的附加关系, 使得网络中的非线性数增加了一倍. 除此之外, 这也缓解了 3D 卷积网络难以优化的问题.

上述方法中, 2D 网络只能获得当前输入帧的空间信息, 缺乏帧间相关信息; 3D 网络虽然可以提取帧间信息, 但其网络参数庞大导致昂贵的运算代价和难以优化的问题. 基于此, 本文提出时间与通道混洗模块, 通过矩阵操作将独立的输入帧转换为具有关联性的全新帧并提取特征信息. 该模块能嵌入主流 2D 卷积网络中构建纯 2D 网络, 在保证参数量的前提下, 实现时空信息的精确提取.

1 时间与通道混洗网络的整体架构

本节首先介绍 ResNet50[12]的构成 block—Bottleneck 并指出其存在的问题, 接着介绍时间与通道混洗模块嵌入 Bottleneck 后的网络构成 block; 最后给出本文网络的整体架构. 时间与通道混洗模块将在第 3 节详细介绍.

1.1 ResNet50 基础 block 问题分析

ResNet50 由 4 个网络层(stage1-4)构成, 每层构成 block 是 Bottleneck, 其结构过程如图 1 所示. 该模块首先对输入特征图使用 1 × 1 卷积核进行卷积操作, 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 卷积核来提取更细化特征. 最后使用 1 × 1 卷积核进行升维操作, 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.

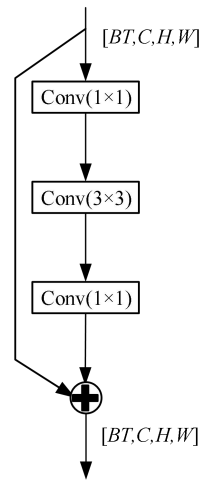


图 1 Bottleneck 结构示意图

1.2 改进后网络构成 block

从上小节分析可知, Bottleneck 只能获得单帧输入的特征信息, 为了获得帧间信息, 在 Bottleneck 中设计并引入时间与通道混洗模块(Temporal-Channel Shuffle Module, TCSM)并命名为 TCSM-Block, 该 block 如图 2 所示.

在 TCSM-Block 中, 输入经过第一个卷积操作后, 对输出进行分支操作, 第一支依旧执行 bottleneck 操作; 第二支先进行 Reshape 操作将维度[BT, C, H, W]分解为[B, C, T, H, W], 接着将输

出送入时间与通道混洗模块进行帧间信息提取，然后将输出的特征维度转换回[BT, C, H, W]. 为了匹配特征维度，对时间与通道混洗模块的输出使用 3 × 3 卷积核(绿色框)进行卷积操作；为了不引入额外的计算参数，该卷积核与 bottleneck 中的 3 × 3 卷积核(蓝色框)实行参数共享. 最后采用元素相加操作实现两支信息的融合，剩下的操作和 bottleneck 完全一致.

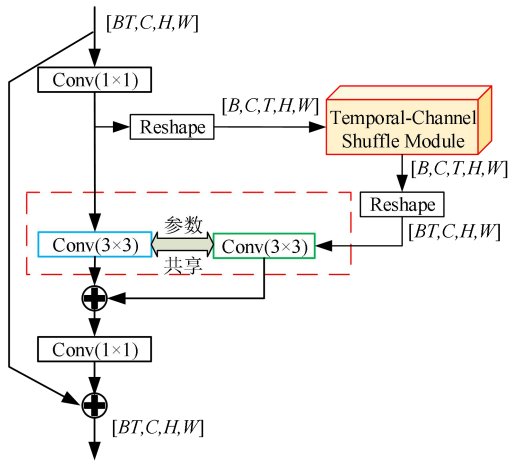


图 2　TCSM-Block 结构示意图

1.3　本文网络整体架构

本文的网络整体结构如图 3 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，红色立方体表示 TCSM-Block，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 TCSM-Block 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. 网络的损失函数采用交叉熵损失函数，其表达式如下：

$$Loss = -\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{1}$$

式(1)中 y_i 表示真实标签，$\hat{y}_i$ 表示预测值，n 表示总类别数.

2　时间与通道混洗模块

本节首先介绍通道混洗模块并指出其问题，然后主要介绍本文提出的时间与通道混洗模块.

2.1　通道混洗模块

在神经网络中，常规 3 × 3 卷积操作往往有着大量的计算参数，为了减少运算量，分组卷积(Group-convolution)逐渐流行起来. 例如，将 Bottleneck 中的常规 3 × 3 卷积操作替换为分组卷积，网络参数将大大减少.

分组卷积简单流程如图 4(a)所示. 首先将输入特征分成 N 组(图中 N=3)，然后在每一组中对特征进行卷积操作，最后将每组的输出级联起来作为整体输出，并送入后续网络. 分组卷积虽然降低了运算量，却将原本的大特征分割成了独立的 N 个小特征，这导致特征间的相关性被破坏.

为了解决这个问题，Zhang 等[13]提出了通道混洗模块(Channel Shuffle)，该过程如图 4(b)所示. 在图 4(b)中，特征首先经过一个分组卷积(GConv1)后分成 3 组，接着将每组特征再细分为 3 个子组，然后将每个子组合并成新的特征组并传入第二个分组卷积(GConv2)中. 对于第二个分组卷积，其输入不再是单一的特征，而是混合后的特征，这使得输出特征既包含原先的特征信息，同时还包含与其他特征的相关信息.

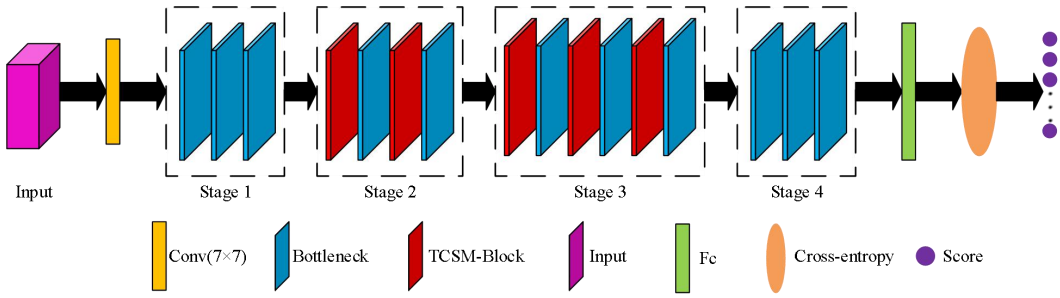
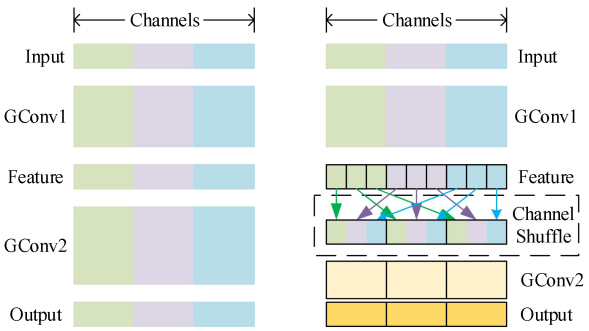


图 3　网络整体架构示意图



a. 分组卷积 b. 通道混洗

图 4 分组卷积中的通道混洗示意图

2.2 时间与通道混洗模块

通道混洗模块有 2 个主要局限: 1)只适用于分组卷积; 2)只能对单张图片输入操作. 视频行为识别通常采用多帧输入, 为了能在纯 2D 卷积网络中提取帧间信息, 受通道混洗模块启发, 本文设计了时间与通道混洗模块(Temporal-Channel Shuffle Module, TCSM). 该模块适用于任意类型的卷积方式, 同时还能用于多帧输入下的情况. 该模块主要流程如图 5 所示.

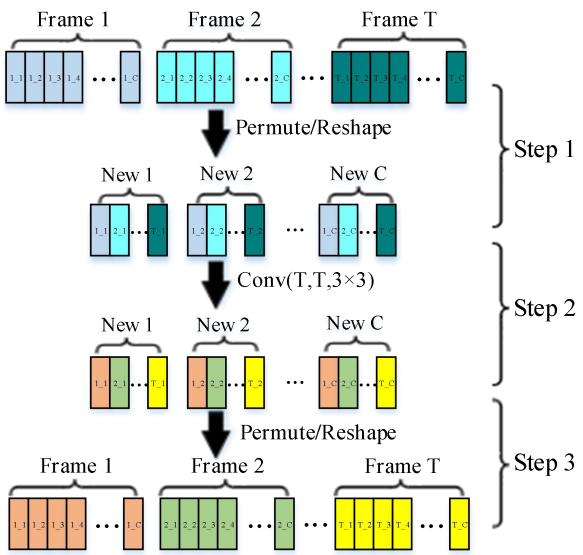


图 5 时间与通道混洗模块

假设网络输入为 T 帧 RGB 图像, 经过多个卷积层后, 每帧图像具有 C 个特征图(通道), 将每个特征图命名为 t_i, 其中 $t\in[1, T]$, $i\in[1, C]$. 例如, 第二帧图像的每个特征图命名为: 2_1, 2_2…2_C. 如图 5 所示, TCSM 的主要流程共分为 3 个步骤, 其中颜色相同表示特征图属于同一帧图像, 并不意味着内容信息相同.

Step1: 2D 卷积只能在单帧图像的所有特征图上进行操作, 因此想要获得帧间相关信息, 首先得构建具有多帧信息的全新帧. TCSM 首先利用矩阵维度转换(Permute/Reshape)操作, 将 T 帧输入中每一帧的第 i 个通道提取出来, 然后合并成新的输入帧. 通过此操作, 原本的 T 帧 C 通道输入就演变成了全新的 C 帧 T 通道输入. 此时的每一帧输入均包含所有帧的原始信息.

Step2: 对得到的新帧进行卷积操作来提取时空关联信息. 此时, 每个新帧的通道数为 T, 卷积操作的输出通道数 T' 需要分析与探讨. 为了引入较少的额外运算量, 从浮点运算数(FLOPs)角度对 T' 进行考虑. 以文献[14]提出的内存访问成本(Memory Access Cost)作为计算损耗, 其表达式如下:

$$S = HWT + HWT' + K^2TT' \tag{2}$$

式(2)中, H 和 W 分别表示特征图的高和宽, K 表示卷积核大小, T 表示输入通道数, T' 表示输出通道数. 对于 $K\times K$ 卷积操作, 其 FLOPs 计算公式如下:

$$M = K^2HWTT' \tag{3}$$

对公式(2)运用均值不等式 $a+b\geq 2\sqrt{ab}$, 可以得到不等式(4):

$$S \geq 2\sqrt{HWT \bullet HWT'} + K^2TT' \tag{4}$$

将公式(3)引入不等式(4), 可以得到以下式子:

$$S \geq 2\sqrt{\frac{HWM}{K^2}} + \frac{M}{HW} \tag{5}$$

由公式(5)可以知道, S 有一个由 M 给出的下限. 根据均值不等式取等条件, 当且仅当 T= T' 时, S 能取到下限. 因此, T= T' 时, 模块引入的额外计算量最少. 为了方便, 文中统一使用 T. 经过卷积, 每个全新帧的每个特征图均具有原始所有帧的相关信息.

Step3: 同样利用矩阵维度转换操作, 将当前的 C 帧 T 通道输入转换为原先的 T 帧 C 通道输入. 相比于最原始的 T 帧 C 通道输入, 此时的 T 帧中的每一通道都含有其他帧的相关信息, 实现了在纯 2D 网络中提取帧间相关信息的目的.

3 实验与分析

本节首先介绍实验数据集, 然后阐述 Baseline 设置与实验设置, 最后对具体实验结果进行分析.

3.1 数据集介绍

本文在最常见的行为识别数据集 UCF101 和 HMDB51 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，类似的视角等.

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.

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. 网络对每帧输入作出预测，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，采用 mini-Kinetics 作为预训练数据集，训练采用小批量随机梯度下降法，动量为 0.9，权值在第 15、25、40 个 epoch 时衰减一次，衰减率为 0.1，总训练数设置为 55. 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.

3.3　实验结果与分析

3.3.1　参数比较

在 2.2 小节中提到, TCSM 中的卷积操作引入的额外参数是少量的，对此，表 1 给出了网络在增加 TCSM-Block 后参数的具体增加值. 当在 ResNet50 的每个 bottleneck 中加入 TCSM，网络只增加了 0.01M 额外参数. 若是在 stage2、stage3 中的部分 bottleneck 加入 TCSM(图 3 所示架构)，网络只增加了 0.002M 额外参数. 由此可见，本文的 TCSM 模块是低内存的、存储友好的.

表 1　参数比较

网络结构	参数量(M)
ResNet50	25.557
ResNet50+TCSM(part-block)	25.559
ResNet50+TCSM(all-block)	25.566

3.3.2　TCSM 位置的影响

对于网络来说, TCSM 是嵌入性模块，那么嵌入位置对于网络的影响就显得很重要. 为了研究嵌入位置对网络的具体影响，表 2 给出了 2 种不同嵌入位置后网络的正确率. Part-block-1 表示在 Baseline 的 stage2、stage3 中的每个 bottleneck 中添加 TCSM; Part-block-2 则是图 2 所示的网络结构.

从结果来看，将 TCSM 嵌入 baseline 后，网络识别正确率得到提升. 但过多位置嵌入 TCSM，提升效果有所下降. 造成这一现象的原因是：对于一个行为，其独特的特征是有限的，因此过多的模块迫使网络学习一些不鲁棒的特征. 此外，由于在测试过程中会接触到不同阶段的特征，过多的嵌入模块会导致特征向量冗余和稀释显著特征.

Part-block-2 表现最优异，因此在后续与其他方法结果进行比较时，均使用此架构下的结果.

表 2　TCSM 位置对网络的影响

方法	主干网络	UCF101(%)
Baseline	ResNet50	89.06
Part-block-1	ResNet50	89.43
Part-block-2	ResNet50	90.59

3.3.3　与主流方法比较

在本小节中，通过具体实验进一步展示了所提出的时间与通道混洗网络与最先进的动作识别方法的比较结果. 关于 UCF101 和 HMDB51 的相关结果分别见表 3 和表 4.

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]、Key-volume[16]; 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[17]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.

本文提出的 TCSM-Net 能够通过只需要多帧 RGB 图像作为输入来探索帧间信息. 表 3 给出了 TCSM-Net 与最先进的动作识别方法的性能比较，为了进行公平比较，这些方法仅使用多帧 RGB 图像作为输入.

从表 3 可以观察到，在 UCF101 数据集上，在所有比较的方法中, TCSM-Net 获得了最好的性能，准确率为 90.6%.

表 3 中的 2D 方法，如 TSN[5]、Slow-fusion[18]等，只能提取单帧 RGB 图像的表观信息，无法获得帧间相关信息; TCSM-Net 则凭借时间与通道混洗

表 3 与主流方法的比较(仅 RGB 输入)

方法	卷积维度(dim)	UCF101(%)	HMDB51(%)
Slow-fusion[18]	2D	65.4	-
C3D[19]	3D	44.0	43.9
LTC[20]	3D	59.9	-
Twostream[4]	2D	73.0	40.5
Twostream+LSTM[21]	2D	82.6	47.1
TSN[5]	2D	85.7	54.6
Key-volume mining[16]	2D	84.5	-
ST-ResNet[22]	2D	82.2	43.4
TLE[15]	3D	86.9	63.2
I3D[17]	3D	84.5	49.8
P3D ResNet[9]	2D	88.6	-
Ours(TCSM-Net)	2D	90.6	63.3

表 4 与主流方法的比较(多流输入)

方法	输入	UCF101(%)	HMDB51(%)
C3D+IDT[19]	RGB+Flow	90.4	-
R(2+1)D[23]	RGB+Flow	95.0	72.7
LTC[20]	RGB+Flow	91.7	64.8
Twostream+LSTM[21]	RGB+Flow	88.6	-
TLE[15]	RGB+Flow	95.4	71.1
TSN[5]	RGB+Flow+RGB_Diff	94.2	69.4
MiCT[10]	RGB+Flow	94.7	70.5
ABM[24]	RGB+Flow	95.1	72.1
Hid-twostream[25]	RGB+Flow	93.2	66.8
STINP[26]	RGB+Flow	94.4	69.6
Ours(TCSM-Net)	RGB+Flow	95.6	75.4

模块提取到了帧间信息，因此 TCSM-Net 可以获得更高的正确率.

表 3 中的 3D 方法，如 C3D[19]、I3D[10]等，也可以获得帧间信息，但效果却不如 TCSM-Net，主要原因有 2 个：1) 3D 网络需要大量数据训练，但 UCF101 数据集数据量并不是很多，这导致这些方法可能出现欠拟合；2) UCF101 数据集每类视频的背景相似且时序性并不强，这使得 3D 方法并不能达到理想结果.

在 HMDB51 数据集上，在所有比较的方法中，TCSM-Net 凭借 63.3%的准确率排在所有方法的第一名. 两个数据集的结果表明本文提出的 TCSM 模块能给基础网络带来较大提升.

在许多先前的论文中，额外的运动信息已经被证明对动作识别有帮助. 基于双流的方法明确地使用运动特征，例如光流，来提高性能. 例如，在 Twostream+LSTM[21]中，单独使用 RGB 作为输入，其在 UCF101 上的正确率只有 82.6%，当引入 flow 作为额外的补充信息时，在 UCF101 上的正确率可以达到 88.6%(详见表 4). 对于 TSN[5]，仅考虑 RGB 输入时，在 UCF101 上只能达到 85.7%的正确率，若是采用多模态(光流、RGB 差分)作为额外信息，TSN 则可以获得 94.2%的正确率(详见表 4).

为了更好地证明 TCSM 的有效性和泛化性，在实验环节设计一个简单的双流 TCSM-Net，其中一个流将 RGB 帧作为输入，另一个流将光流作为输入. RGB 流结构和前文一致,对于光流分支，除了将第一个卷积层的通道尺寸从 3 扩大到 20 之外，其余结构均与 RGB 流相同. 通道尺寸扩大的原因是对于光流分支，TCSM-Net 采样 10 个光流图像并堆叠在一起作为输入，每个光流图像由两个通道(水平方向和垂直方向)组成. 我们在训练期间分别优化 RGB 流和光流，并在测试期间简单地将这两个流的推断结果做平均然后作为最终预测.

从表 4 结果来看，无论是在 UCF101 上，还是在 HMDB51 上，TCSM-Net 的正确率均排在第一名.

这证明 TCSM 在光流输入上也能带来效果提升，同时再次验证 TCSM 模块的有效性.

4 结 语

本文提出了时间与通道混洗模块下的人体行为识别方法. 通过分析现有通道混洗模块的局限性，提出了时间与通道混洗模块. 为了验证模块的有效性，分别从额外网络参数、嵌入位置、精度提升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间与通道混洗模块的有效性.

参考文献(References):

[1] Ikizier-cinbis N, Sclaroff S. Object, sence and actions: Combining multiple features for human action recognition[C]//European Conference on Computer Vision, Heraklion, Crete, Greece, 2010, 6311: 494-507

[2] Zhang Liang, Lu Mengmeng, Jiang Hua. 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(in Chinese)
(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, 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, 2014: 568-576

[5] Wang L, Xiong Y, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition[C] //Proceedings of European Conference on Computer Vision, Springer, Cham, 2016: 20-36

[6] Zhang Bingbing, Ge Shuyu, Wang Qilong, *et al*. Multi-order Information Fusion Method for Human Action Recognition[J/OL]. Acta Automatica Sinica, https://doi.org/10.16383/j.aas.c180265(in Chinese)
(张冰冰，葛疏雨，王旗龙，等. 基于多阶信息融合的行为识别方法研究[J/OL]. 自动化学报，https://doi.org/10.16383/j.aas.c180265)

[7] Song Lifei, Weng Liguo, Wang Lingfeng, *et al*. Multi-scale 3D Convolution Fusion Two-Stream Networks for Action Recognition[J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(11):99-108(in Chinese)
(宋立飞，翁理国，汪凌峰，等. 多尺度输入 3D 卷积融合双流模型的行为识别方法[J]. 计算机辅助设计与图形学学报，2018, 30(11):99-108)

[8] Ji S, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231

[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, 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, 2018: 449-458

[11] 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, 2018: 6450-6459

[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, USA, 2016: 770-778

[13] Zhang X, Zhou X, Lin M, *et al*. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, 2018: 6848-6856

[14] Ma N, Zhang X, Zheng H T, *et al*. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European Conference on Computer Vision, Munich, Germany, 2018: 116-131

[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, 2017: 2329-2338

[16] W. Zhu, J. Hu, G. Sun, *et al*. A key volume mining deep framework for action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 1991–1999

[17] 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, 2017: 6299-6308

[18] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colombus, USA, 2014: 1725-1732

[19] 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, 2015: 4489-4497

[20] G. Varol, I. Laptev, C. Schmid. Long-term temporal convolutions for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017

[21] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, *et al.* Beyond short snippets: Deep networks for video classification[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 4694-4702

[22] Feichtenhofer C, Pinz A, Wildes R. Spatiotemporal residual networks for video action recognition[C]//Advances in Neural Information Processing Systems, Barcelona, Spain, 2016:

3468-3476

[23] 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, 2018: 6450-6459

[24] Zhu X, Xu C, Hui L, *et al*. Approximated bilinear modules for temporal modeling[C]//Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 2019: 3494-3503

[25] Zhu Y, Lan Z, Newsam S, *et al*. Hidden two-stream convolutional networks for action recognition[C]//Asian Conference on Computer Vision, Springer, Cham, 2018: 363-378

[26] Chen J, Kong J, Sun H, *et al*. Spatiotemporal Interaction Residual Networks with Pseudo3D for Video Action Recognition[J]. Sensors, 2020, 20(11):3126

龚苏明，18252588127@163.com， 15852715278， QQ: 1254849950
陈莹， chenying@jiangnan.edu.cn， 13861855711