时空压缩激励残差乘法网络的视频动作识别

罗会兰，童康
（江西理工大学信息工程学院，江西 赣州 341000）

摘 要：针对双流网络结构中浅层网络和一般深度模型学习空间信息和时间信息的不足，提出将压缩激励残差网络用于空间流和时间流的动作识别，同时将恒等映射核作为时间滤波器注入网络中捕获长期时间依赖性。为了进一步加强压缩激励残差网络的空间信息和时间信息之间的交互，采用时空特征相乘融合，并研究空间流和时间流乘法融合方式、次数以及位置对识别性能的影响。鉴于单个模型获得性能的局限性，提出了 3 种不同的策略生成多个模型，并使用直接平均与加权平均集成以得到最终识别结果。HMDB51 和 UCF101 数据集上的实验结果表明，所提时空压缩激励残差乘法网络能够有效提升动作识别性能。

关键词：动作识别；时空流；压缩激励残差网络；相乘融合；多模型集成
中图分类号：TP391
文献标识码：A
doi: 10.11959/j.issn.1000−436x.2019194

Spatiotemporal squeeze-and-excitation residual multiplier network for video action recognition

LUO Huilan, TONG Kang
School of Information Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China

Abstract: Aiming at the shortcomings of shallow networks and general deep models in two-stream network structure, which could not effectively learn spatial and temporal information, a squeeze-and-excitation residual network was proposed for action recognition with a spatial stream and a temporal stream. Meanwhile, the long-term temporal dependence was captured by injecting the identity mapping kernel into the network as a temporal filter. Spatiotemporal feature multiplication fusion was used to further enhance the interaction between spatial information and temporal information of squeeze-and-excitation residual networks. Simultaneously, the influence of spatial-temporal stream multiplication fusion methods, times and locations on the performance of action recognition was studied. Given the limitations of performance achieved by a single model, three different strategies were proposed to generate multiple models, and the final recognition result was obtained by integrating these models through averaging and weighted averaging. The experimental results on the HMDB51 and UCF101 datasets show that the proposed spatiotemporal squeeze-and-excitation residual multiplier networks can effectively improve the performance of action recognition.

Key words: action recognition, spatiotemporal stream, squeeze-and-excitation residual network, multiplication fusion, multi-model ensemble

1 引言

视频作为信息的主要载体之一，已越来越多地被人们共享。如何理解和分析这些海量涌现的视频数据至关重要。视频中的人体动作识别[1-4]一直广受研究者的青睐，在日常生活安全、视频信息检索、

收稿日期：2019−03−04；**修回日期**：2019−07−17
基金项目：国家自然科学基金资助项目（No.61862031）；江西省自然科学基金资助项目（No.20171BAB202014）；江西省赣州市“科技创新人才计划”基金资助项目
Foundation Items: The National Natural Science Foundation of China(No.61862031), Jiangxi Natural Science Foundation(No.20171BAB202014), “Science and Technology Innovation Talent Plan” Project of Ganzhou, Jiangxi Province

公共视频监控、人机交互等领域都有广泛的应用。当前视频中的人体动作识别研究方法大致可以分为 2 类：传统手动特征提取方法和基于深度学习的方法。

传统手动特征提取方法是将特征的提取与后续动作识别的训练分成 2 个独立的过程，在获得动作视频的特征表示后输入机器学习算法进行训练，实现最终的分类与识别。比较有代表性的早期工作有 Bobick 等[5-6]采用运动能量图像和运动历史图像来解释图像序列中人的运动。Yilmaz 等[7]提出通过在时间轴上叠加目标的轮廓来构建时空卷，再根据时空卷的不同属性来识别动作。该类方法需要将运动人体从背景中分割出来，所以在复杂动态背景情况下效果不好。Wang 等[8]提出了利用时空兴趣点（STIP, space time interest point）来描述视频，STIP 特征是利用角点探测器获得兴趣点进行跟踪并提取描述符信息。Klaser 等[9]通过采样和跟踪多个尺度上每帧的稠密点来提取稠密轨迹（DT, dense trajectory），并用 DT 表示视频。DT 特征是对视频进行稠密采样，捕捉运动轨迹，并沿着光流方向提取轨迹的方向梯度直方图、光流直方图和运动边界直方图这些描述符信息。Wang 等[10]提出了改进的稠密轨迹（IDT, improving dense trajectory），对人物进行了框定，消除了相机抖动及背景杂乱的影响。基于IDT特征的动作识别方法获得的识别准确率一度达到世界领先水平。

不同于传统手动特征提取方法，基于深度学习的方法旨在自动从视频中学习到有效特征用于动作识别。为了便于处理视频，Ji 等[11]提出了三维卷积网络，并将其用于识别视频中的人类动作。在此基础上，Du 等[12]提出了深度三维卷积神经网络，该方法直接利用深度三维卷积网络中的三维卷积和三维池化对 RGB 视频进行处理，并利用大规模有监督视频数据集进行训练获得 C3D（convolutional 3D）模型。后来，Tran 等[13]将三维卷积和残差网络相结合，并在数据集 Sports-1M[14]上训练获得 Res3D（residual 3D）模型，它比 C3D 模型小一半且运行速度更快。为了更好地获得时间信息和空间信息，Simonyan 等[15]提出了双流卷积神经网络进行动作识别，分别使用 RGB 视频帧和光流图片作为输入进行训练，以构成空间流网络和时间流网络，并用这 2 个网络流的分类得分的平均值作为最终分类结果。在此基础上，很多基于双流卷积神经网络的方法，包括时间分割网络[16]、时空残差网络[17]、动作变换[18]、时空金字塔网络[19]等被提出，并且获得了不错的识别率。针对双流卷积神经网络中时间流和空间流平均融合方法的不足，Feichtenhofer 等[20]提出了在卷积层之后进行融合的新方法，实现了从分类器级融合到特征级融合的转变。随后，Feichtenhofer 等[21]探索了许多连接外观流和运动流的方法，并提出了乘法交互的跨流残差连接，这种新的时空乘法网络结构在视频中的人体动作识别上获得了良好的性能。

有些研究者试图构造更多流的网络来尽可能多地获取到视频中的动作特征信息。Wang 等[22]提出了全局时空三流卷积神经网络结构，利用单帧图像、10 帧光流堆叠以及运动堆叠的差分图像作为卷积神经网络的输入，训练获得空间流、局部时间流和全局时间流特征。对这些学习到的特征先进行 PCA（principal component analysis）-Whitening 操作，然后进行 soft-VLAD（soft vector of locally aggregated descriptor）矢量编码，最后使用支持向量机分类。Bilen 等[23]提出了四流网络结构，分别应用排序池化和近似排序池化对 RGB 图像和光流进行编码得到动态图像，并将其输入卷积神经网络训练得到 RGB 动态图像流网络和动态光流网络，结合原始 RGB 流网络和光流网络形成四流网络结构，最后用四流网络输出得分的均值来预测动作类，获得了不错的识别效果。

本文基于双流卷积神经网络结构，提出了一种用于动作识别的时空压缩激励残差乘法网络。受残差网络模型[24]和压缩激励（SZ, squeeze and excitation）网络模型[25]的启发，本文将压缩激励块和残差网络模型结合的压缩激励残差网络模型用于空间流和时间流。受文献[21]中的时空乘法交互以及恒等映射滤波器的启发，本文对空间压缩激励残差网络模型和时间压缩激励残差网络模型采用特征相乘融合，以更好地学习时空特征；同时，将恒等映射核作为时间滤波器注入网络模型中，以此来学习长期时间依赖关系。鉴于单个模型获得性能的局限性以及受集成学习思想的启发，本文使用 3 种不同的策略生成多个模型，并对它们进行均值及加权平均集成方法来获得最终的识别结果。

本文的贡献介绍如下。1) 将图像识别领域的残差网络和压缩激励网络结合的压缩激励残差网络迁移到视频动作识别中；2) 以 RGB 和光流图

片为输入，训练获得双流卷积神经网络，同时注入时间滤波器对空间流和时间流进行特征级别的乘法融合；3) 采用集成学习思想，将不同策略获得的多个模型进行直接平均和加权平均集成；4) 进行了一系列比较分析实验，结果表明本文通过特征级别乘法融合以及多模型集成获得了很好的识别效果。

2　技术方法

本文动作识别的整体框架结构如图 1 所示。首先，将压缩激励块和残差网络结合的压缩激励残差网络模型作为网络的基础模型，同时注入时间滤波。然后，用 RGB 视频帧和光流数据分别进行训练，获得空间流网络模型和时间流网络模型；在此基础上，将空间流网络训练获得的空间压缩激励残差网络模型与时间流网络训练获得的时间压缩激励残差网络模型进行乘法融合并再次训练。最后，利用不同策略训练获得多个时空压缩激励残差乘法网络模型，通过直接平均和加权平均对这些模型进行集成以获得最终的识别结果。

2.1　压缩激励块

压缩激励块的原理如图 2 所示。任何一个卷积层的输出都可以通过压缩激励块实现跨通道全局信息依赖关系的学习，每个通道得到一个尺度系数。由图 2 可知，对于一个输出维度为 $W\times H\times C$ 的卷积层，首先通过全局平均池化获得维度为 $1\times1\times C$ 的输出，得到每个特征通道的全局信息；然后通过 2 个全连接层来学习不同特征通道间的依赖关系，2 个全连接层后面分别采用了 ReLU 和 Sigmoid 激活函数对全连接层的输出激活；最后将压缩激励块得到的输出 $1\times1\times C$ 和最初卷积层的输出 $W\times H\times C$ 相乘，即每个特征通道乘以一个学习得到的尺度系数。一个输出维度为 $W\times H\times C$ 的卷积层通过压缩激励块操作之后，特征维度没有发生改变，但充分学习了 C 个通道间的全局依赖关系。

2.2　压缩激励残差网络

残差网络对学习深度表征十分有效，而压缩激励（SE, squeeze-and-excitation）块具有学习跨通道全局依赖的能力，它可以嵌入任何的卷积层后，所以本文将压缩激励块集成到残差网络中构建压缩激励残差网络。图 3 为本文构建的一个 50 层的压缩激励残差网络结构，图中省略了跳跃连接。

2.3　恒等特征的时间滤波

一维卷积可以有效捕捉时间依赖性。为了提供更大间隔的时间支持，本文使用一维时间卷积，它结合了初始化为恒等映射的特征空间变换，可以将时间滤波加入残差层中，从而产生局部影响，也可以加入跳跃连接层，产生全局影响。

第 l 层后的时间滤波操作如式(1)所示，跨越 C_l 个特征通道实现时间滤波。

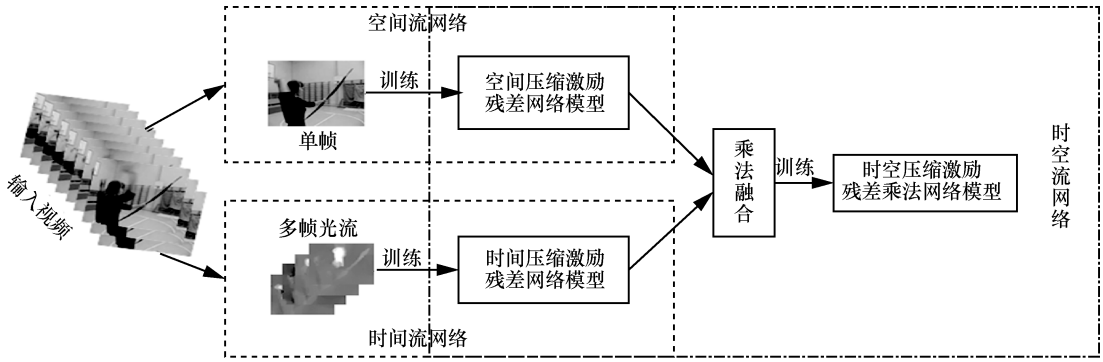

图 1　动作识别整体框架结构

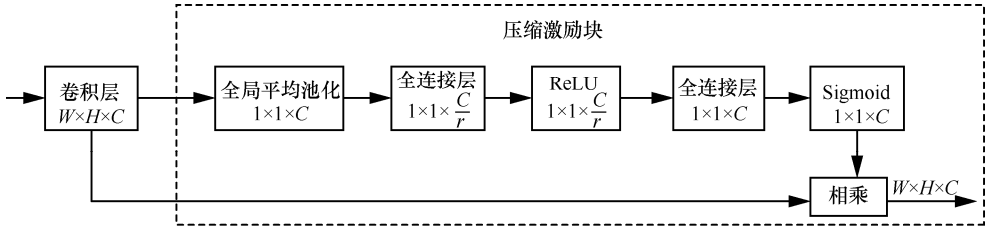

图 2　压缩激励块的原理

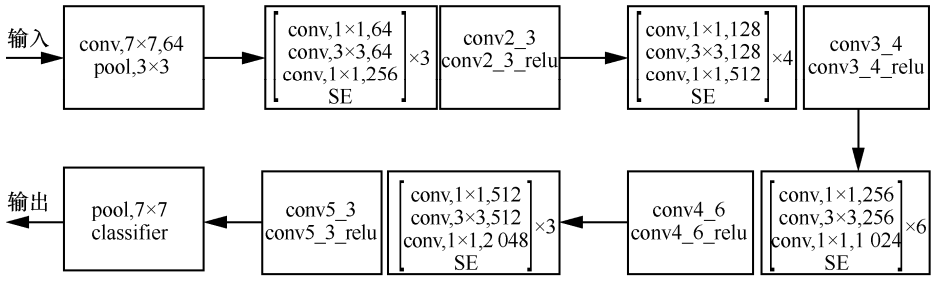

图 3　压缩激励残差网络结构

$$x_{l+1} = x_l * \hat{W}_l + b_l \tag{1}$$

其中，x_{l+1} 和 x_l 分别表示第 $l+1$ 层和第 l 层；*为卷积操作；偏置 b_l 初始化为 0；$\hat{W}_l \in R^{1\times1\times T\times C_l\times C_l}$ 是跨越时间为 $t=1,\cdots,T$ 的时间滤波器权值，它由特征通道间堆叠的恒等映射 $I \in R^{1\times1\times1\times C_l\times C_l}$ 来初始化。时间滤波器权值计算式如式(2)所示。

$$\hat{W}_l = I \otimes f \tag{2}$$

其中，$\otimes$ 表示张量外积，f 表示一个长度为 T 的一维时间滤波器。

在时间滤波的基础上，同时引入全局时间池化，放置于最后一个卷积层，用于捕获全局时间信息，在时间范围 $1\leqslant t\leqslant T$ 内，给定 $x(i,j,t,c)$，全局最大时间池化计算式如式(3)所示。

$$x(i,j,c) = \max_{1\leqslant t\leqslant T} x(i,j,t,c) \tag{3}$$

2.4　时间流和空间流的乘法融合

为了更好地学习时空流网络特征，本文采用特征级别的融合方法，即对空间压缩激励残差网络和时间压缩激励残差网络进行乘法交互。2 个压缩激励残差块双向连接（时间流到空间流、空间流到时间流）乘法融合如图 4 所示。将时间流压缩激励残差块的输出与空间流对应的压缩激励残差块输出进行元素级别的乘法，这样通过空间流残差单元的信息就被时间信号所调整。类似地，将空间流压缩激励残差块的输出与时间流的压缩激励残差块的输出进行乘法融合，时间流的信息被空间信号所调整。通过时间流与空间流的乘法融合，学习到特征级别的时空信息，有助于区分外观上相似的动作。

本文提出的注入时间滤波器的压缩激励残差乘法网络结构如图 5 所示。其中，⊙表示乘法融合交互点，inject 表示注入时间滤波器。在图 5 所示的结构中，空间流与时间流的乘法融合交互分别在 conv2_x、conv3_x、conv4_x、conv5_x 中的第一个压缩激励残差块后进行，inject 除了在 conv5_x 的最后一个压缩激励残差块后进行，还分别在 conv2_x、conv3_x、conv4_x、conv5_x 的第二个压缩激励残差块中的 3×3 卷积后进行。

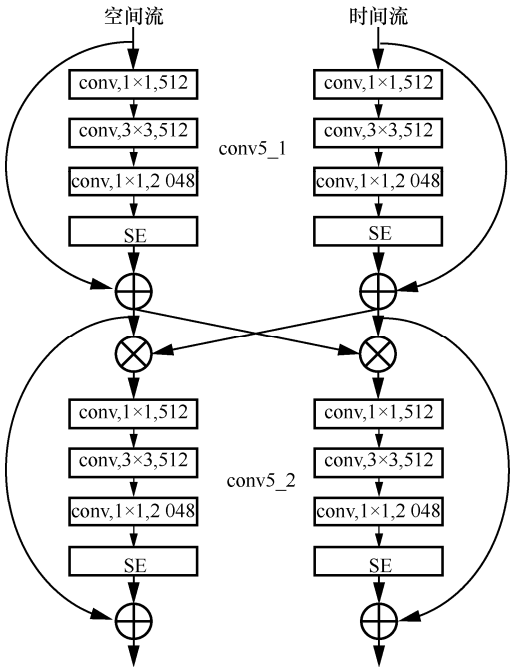

图 4　乘法融合示意

2.5　多模型集成

由于空间流与时间流的乘法融合方式（空间流到时间流、时间流到空间流）、次数和位置可以变化，以及受集成学习思想的启发，对本文所提的注入时间滤波器的压缩激励残差乘法网络架构，采用不同的乘法融合策略，在不同的训练数据划分子集上学习，从而可以获得多个动作分类模型，并在测试阶段对这些分类模型结果进行集成以进一步提升识别效果。

针对不同策略生成的多个模型，本文采用直接

layer	conv1	pool1	conv2_*x*	conv3_*x*	conv4_*x*	conv5_*x*	pool5
block	conv 3×3 64	max pool 3×3	[conv,1×1,64 conv,3×3,64 conv,1×1,128 SE] ⊙ [conv,1×1,64 conv,3×3,64 inject conv,1×1,256 SE] [conv,1×1,64 conv,3×3,64 conv,1×1,256 SE]	[conv,1×1,128 conv,3×3,128 conv,1×1,256 SE] ⊙ [conv,1×1,128 conv,3×3,128 inject conv,1×1,512 SE] [conv,1×1,128 conv,3×3,128 conv,1×1,512 SE]×2	[conv,1×1,256 conv,3×3,256 conv,1×1,1 024 SE] ⊙ [conv,1×1,256 conv,3×3,256 inject conv,1×1,1 024 SE] [conv,1×1,256 conv,3×3,256 conv,1×1,1 024 SE]×4	[conv,1×1,512 conv,3×3,512 conv,1×1,2 048 SE] ⊙ [conv,1×1,512 conv,3×3,512 inject conv,1×1,2 048 SE] [conv,1×1,512 conv,3×3,512 conv,1×1,2 048 SE]	7×7 max

图 5　注入时间滤波器的压缩激励残差乘法网络结构

平均法和加权平均法进行集成。直接平均法就是对不同模型产生的类别置信度求均值得到最终的测试结果。而加权平均法则是在直接平均方法基础上，通过加入权重来调节不同模型输出间的重要程度。假设共有 N 个模型待集成，对测试样本 D，其测试结果为 N 个 M 维（M 为数据的标记空间大小）向量 $\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_N$。直接平均法和加权平均法对应的计算式分别如式(4)和式(5)所示。

$$\text{Score}=\frac{\sum_{i=1}^{N}\boldsymbol{q}_i}{N} \tag{4}$$

$$\text{Score}=\frac{\sum_{i=1}^{N}w_i\boldsymbol{q}_i}{N} \tag{5}$$

其中，w_i 对应第 i 个模型的权重，$w_i\geqslant 0$ 且 $\sum_{i=1}^{N}w_i=1$。

3　实验结果

本文实验采用 Matlab 2017a 的仿真软件以及 MatConvNet 的深度学习工具包。实验环境配置如下：操作系统为 64 位的 Windows10，CPU 为 Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60 GHz；内存为 512 GB，显卡为 16 GB 的 NVIDIA Tesla P100-PCIE。

3.1　实验数据集及实验设置

UCF101[26]数据集是最流行的动作识别数据集之一，包含 13 320 个来自 101 个动作类别的视频片断。其中，每一个类别至少有 100 个视频片段，每一个片段持续 3～10 s。该数据集的 101 个类别可以分为五大类，包括体育运动、乐器演奏、人与人之间的交互、身体运动、人与对象的交互。由于该数据集来源于现实环境，包含杂乱背景、相机抖动、遮挡、不同光照条件等各种因素的影响，故该数据集具有一定的挑战性。

HMDB51[27]数据集是一个大而真实的视频集合，包含 51 个动作类别，涵盖了 6 766 个视频片断。这些视频片断主要来源于电影，只有一小部分来自公共数据库，并且每一个片断都包含一个人类活动。该数据集的行为类别包括普通面部动作、操纵对象面部动作、一般身体运动、与对象交互运动、与人交互运动共 5 种类型。HMDB51 数据集来源不同，并伴有遮挡、相机移动、复杂背景、光照条件变化等诸多因素的影响，相较于 UCF101 数据集更具挑战性。

本文采用交叉验证方法进行训练，UCF101 数据集的训练集的 3 种不同划分分别为 split_1、split_2 和 split_3。每种数据划分将全部训练视频数据按 7:3 的比例分为训练集和验证集。具体步骤如下。每个动作类共有 25 组训练视频，其中 split_1 将前面 7 组视频作为验证集，剩下的 18 组视频作为训练集；split_2 将第 8 组到第 14 组视频作为验证集，其余的作为训练集；split_3 则将第 15 组到第 21 组视频作为验证集，剩下的视频作为训练集。HMDB51 数据集的 3 种不同划分和 UCF101 数据集类似。除了从视频中提取 RGB 视频帧外，还预先计算光流并以 JPEG 形式存储。本文采用文献[21]中使用的 UCF101、HMDB51 的视频帧和光流数据。

本文使用 50 层的压缩激励残差网络作为基础模型，并将其分别用于空间流网络和时间流网络。空间流网络和时间流网络的训练是分开进行的，并且都使用动量为 0.9 的随机梯度下降。时间流网络使用 10 帧堆叠的光流帧作为输入，通过对光流的中间和边缘的随机裁剪，将图片大小调整为 224 像

素×224 像素。时间流网络训练的批量处理大小为 128 张图片，初始学习率为 0.01，每次减小到原来的$\frac{1}{10}$，直至减小到初始值的$\frac{1}{1\,000}$。空间流网络以大小为 224 像素×224 像素的 RGB 图片作为输入，批量处理大小为 256 张图片，学习率分别为 0.01、0.001、0.000 1。

3.2　实验及分析

3.2.1　单流网络性能分析

本节评估训练获得的空间压缩激励残差网络模型和时间压缩激励残差网络模型在 UCF101 和 HMDB51 数据集上的识别效果，以及它们在不同划分下训练后获得测试性能的差异。表 1 给出了空间流网络和时间流网络在 UCF101 和 HMDB51 数据集上的识别准确率。从表 1 空间流和时间流网络的对比可以看出，时间流网络在 HMDB51 和 UCF101 数据集上的识别准确率都要高于空间流网络。在 UCF101 数据集上，时间流网络识别准确率比空间流网络识别准确率高 2.8%；在 HMDB51 数据集上，时间流网络比空间流网络高 10.9%。数据集 HMDB51 受相机抖动、复杂背景等因素的影响要大于数据集 UCF101，并且数据集 HMDB51 上同一动作具有大的类内散度以及不同动作具有小的类间散度的程度要大于数据集 UCF101。时间流网络能够更好地对这两者的影响进行补偿，这可能就是时间流网络在 HMDB51 数据集上较空间流网络提升比在 UCF101 数据集上高的原因。此外，HMDB51 和 UCF101 数据集在不同划分下训练后获得的测试识别准确率也有差异，在 HMDB51 数据集上，时间流网络和空间流网络均在 $split_2$ 划分下训练后取得最高的测试识别准确率；而在 UCF101 数据集上，空间流网络在 $split_1$ 划分下训练后的测试识别准确率较高，时间流网络却在 $split_3$ 划分下训练后的测试识别准确率最高。这也说明训练数据对识别性能有较大影响。

表 1　HMDB51 和 UCF101 数据集上识别准确率

数据集	划分	识别准确率	
		空间流网络	时间流网络
UCF101	$split_1$	83.40%	84.20%
	$split_2$	83.30%	85.90%
	$split_3$	83.00%	87.80%
	平均值	83.20%	86.00%
HMDB51	$split_1$	47.60%	57.90%
	$split_2$	48.60%	58.70%
	$split_3$	44.70%	57.10%
	平均值	47.00%	57.90%

3.2.2　空间流与时间流乘法融合方式、次数及位置对识别性能的影响分析

本节实验分析了使用相同乘法融合方式情况下，即采用从时间流到空间流的乘法融合方式，融合次数和位置对识别性能的影响，实验结果如表 2 所示，结果报告了在 HMDB51 数据集 $split_1$ 划分下训练后获得的测试识别准确率。其中“conv2_1_relu 和 conv2_1”表示从时间流的 conv2_1 层连接到空间流的 conv2_1_relu 层进行乘法融合，其他依次类推。

从表 2 可以看出，对于单次融合来说，“conv5_1_relu 和 conv5_1”融合获得了 67.1%的识别准确率，它比“conv2_1_relu 和 conv2_1”“conv3_1_relu 和 conv3_1”“conv4_1_relu 和

表 2　HMDB51 数据集上从时间流到空间流在不同次数和位置下融合的识别准确率

融合次数	融合位置	识别准确率
单次融合	conv2_1_relu 和 conv2_1	65.9%
	conv3_1_relu 和 conv3_1	66.1%
	conv4_1_relu 和 conv4_1	66.5%
	conv5_1_relu 和 conv5_1	67.1%
两次融合	conv4_1_relu 和 conv4_1, conv5_1_relu 和 conv5_1	69.7%
三次融合	conv3_1_relu 和 conv3_1, conv4_1_relu 和 conv4_1, conv5_1_relu 和 conv5_1	69.1%
四次融合	conv2_1_relu 和 conv2_1, conv3_1_relu 和 conv3_1, conv4_1_relu 和 conv4_1, conv5_1_relu 和 conv5_1	67.6%

conv4_1”融合分别高了 1.2%、1.0%和 0.6%。由此可见，从时间流 conv5_1 层连接到空间流 conv5_1_relu 层融合效果更好，这可能是由于更高的卷积层学到的特征更完整、更具有判别性。

从表 2 还可以看出，在“单次融合、两次融合、三次融合和四次融合”这些不同次数的融合中，“conv4_1_relu 和 conv4_1&conv5_1_relu 和 conv5_1”两次融合和“conv3_1_relu 和 conv3_1&conv4_1_relu 和 conv4_1&conv5_1_relu 和 conv5_1”三次融合分别取得了最高识别准确率和次高识别准确率，识别率分别为 69.7%和 69.1%。而“conv2_1_relu 和 conv2_1&conv3_1_relu 和 conv3_1&conv4_1_relu 和 conv4_1&conv5_1_relu 和 conv5_1”四次融合的识别准确率比三次融合和两次融合分别低了 1.5%和 2.1%。造成这个差异可能的一个原因是“conv2_1_relu 和 conv2_1”底层卷积层融合学到的更多是颜色、边缘等浅层特征，并没有学到高层具有的判别性语义特征，将底层卷积层和其他相对高层的卷积层融合一定程度上降低了识别的准确率。

同样地，为了探究不同融合方式对识别性能的具体影响，进一步实验分析了在不同融合次数及位置情况下，将融合方式设置成从空间流到时间流的乘法融合，在 HMDB51 训练集第一划分下训练，在 HMDB51 测试集上的识别准确率，如表 3 所示。从表 3 的实验结果可以看出，采用“时间流到空间流”比采用“空间流到时间流”融合的效果更优。在单次融合中，“conv2_1_relu 和 conv2_1”采用“时间流到空间流”融合识别准确率比采用“空间流到时间流”高 1.3%。而“conv5_1_relu 和 conv5_1”采用“时间流到空间流”比采用“空间流到时间流”高 2.1%。在两次融合、三次融合和四次融合中，采用“时间流到空间流”融合的识别准确率比采用“空间流到时间流”融合的识别准确率分别高 7.6%、12.0%和 15.6%。

从表 3 还可以看出，在采用“空间流到时间流”的融合方式时，“conv5_1_relu 和 conv5_1”单次融合取得了最好的识别效果，识别准确率达到 65.0%。而随着融合次数的增加，两次融合、三次融合和四次融合在采用“空间流到时间流”融合时的识别准确率却呈逐渐下降趋势，识别率分别为 62.1%、57.1%和 52.0%。造成识别率下降的原因可能是：相较于空间流网络来说，时间流网络学习能力更强，学到的特征更具判别性；而将学习特征能力相对不太强的空间流网络特征注入时间流网络融合，一定程度上会干扰原本时间流网络对特征的学习，随着融合次数的增多，将可能带来负面影响，从而造成识别率逐渐降低。

比较表 1 和表 3 的结果也可以看出，采用时间流到空间流的融合方式，相较于单个空间流和时间流网络，性能都有了较大的提升。

综合以上对比分析，可以得到如下结论。

1) 单次融合中，在更高层位置融合所获得的识别效果更优。

2) 融合次数为“conv4_1_relu 和 conv4_1&conv5_1_relu 和 conv5_1”的两次融合所获得的

表 3　　HMDB51 数据集上不同融合方式下的识别准确率

融合次数	融合位置	融合方式	识别准确率
单次融合	conv2_1_relu 和 conv2_1	时间流到空间流	65.9%
		空间流到时间流	64.6%
	conv5_1_relu 和 conv5_1	时间流到空间流	67.1%
		空间流到时间流	65.0%
两次融合	conv4_1_relu 和 conv4_1, conv5_1_relu 和 conv5_1	时间流到空间流	69.7%
		空间流到时间流	62.1%
三次融合	conv3_1_relu 和 conv3_1, conv4_1_relu 和 conv4_1, conv5_1_relu 和 conv5_1	时间流到空间流	69.1%
		空间流到时间流	57.1%
四次融合	conv2_1_relu 和 conv2_1, conv3_1_relu 和 conv3_1, conv4_1_relu 和 conv4_1, conv5_1_relu 和 conv5_1	时间流到空间流	67.6%
		空间流到时间流	52.0%

识别效果更优。

3) 采用时间流到空间流的乘法融合方式所获得的识别效果更优。

3.2.3 不同策略下产生的多模型集成对识别性能的影响分析

为了分析生成多个模型的不同策略对集成性能的影响，本节实验比较了 3 种不同的策略。

策略 1 固定融合方式为“时间流到空间流”，分别使用如表 2 所示的“三次融合”和“两次融合”2 种融合模式，分别在数据集 HMDB51 的 3 个划分上训练获得 6 个模型进行集成。

策略 2 固定使用如表 2 所示的“两次融合”，分别使用“时间流到空间流”和“空间流到时间流”2 种融合方式，在数据集 HMDB51 的 3 个划分上训练获得 6 个模型进行集成。

策略 3 受“轮数集成”[28]的启发，固定融合方式为“时间流到空间流”以及使用“两次融合”，在数据集 HMDB51 的 3 个划分上训练，分别取每个划分上训练得到的最后 2 轮模型共获得 6 个模型进行集成。

最后，分别对策略 1、策略 2 和策略 3 生成的 6 个模型结果进行直接平均法和加权平均法集成以得到最终的识别结果。对于加权平均法中权值的设置，根据不同模型在验证集上各自单独的准确率而定，高准确率的模型权值较高，低准确率模型的权值较小。对于策略 1 训练获得的 6 个模型，先将其按验证集上的准确率进行排序，然后将准确率第一和第二的分为一组，准确率第三和第四的分为一组，准确率第五和第六的分为一组。这三组分别称为高准确率组、次高准确率组和低准确率组，权值分别为 0.30、0.15 和 0.05，策略 2 和策略 3 进行相同操作。

表 4 为不同策略下产生的多模型采用直接平均和加权平均集成后，在 HMDB51 数据集上的识别准确率。从表 4 可以看出，采用加权平均法比采用直接平均法在策略 1、策略 2 和策略 3 上分别高 0.7%、2.0%和 0.5%。由此可见，采用加权平均法比采用直接平均法更有利于识别准确率的提升，特别地，策略 3 生成的 6 个模型进行加权平均集成后在 HMDB51 数据集上获得了 69.3%的识别准确率。

表 4 不同策略下产生的多模型集成在 HMDB51 数据集上的识别准确率

方法	识别准确率
策略 1（直接平均）	68.5%
策略 1（加权平均）	69.2%
策略 2（直接平均）	65.6%
策略 2（加权平均）	67.6%
策略 3（直接平均）	68.8%
策略 3（加权平均）	69.3%

3.2.4 和当前其他动作识别算法的性能比较

表 5 为本文方法与当前其他动作识别算法在 UCF101 和 HMDB51 数据集上识别准确率的对比。表 5 中给出的本文方法的结果，是使用生成多个模型的策略 3 以及加权平均的集成方法获得的结果。本文方法在 HMDB51 和 UCF101 数据集上分别获得了 69.3%和 92.4%的识别准确率。从表 5 可以看出，虽然在 UCF101 数据集上本文方法较时空乘法网络、时空金字塔网络识别率要分别低 1.8%和 0.8%，但是相较于改进稠密轨迹方法、三维残差卷积网络、双流卷积神经网络及三流卷积神经网络，本文方法分别获得了 6.0%、6.6%、4.4%和 0.3%的准确率的提升。相较于 UCF101 数据集，本文方法

表 5 HMDB51 和 UCF101 数据集上平均识别准确率

方法	UCF101	HMDB51
改进的稠密轨迹[10]	86.4%	61.7%
三维残差卷积网络[13]	85.8%	54.9%
双流卷积神经网络[15]	88.0%	59.4%
卷积双流网络融合[20]	91.8%	64.6%
时空金字塔网络[19]	93.2%	66.1%
时空乘法网络[21]	94.2%	68.9%
三流卷积神经网络[22]	92.1%	67.2%
语义图像网络[29]	92.1%	65.8%
本文方法（策略 3+加权平均）	92.4%	69.3%

在 HMDB51 数据集上获得了更高程度的性能提升，特别地，相比较于识别率较低的三维残差卷积网络和双流卷积神经网络，本文方法分别获得了 14.4%和 9.9%的准确率提升；相较于识别率较高的时空乘法网络和三流卷积神经网络，本文方法也分别获得了 0.4%和 2.1%的准确率提升。时空乘法网络中的外观流（即空间流）和运动流（即时间流）分别使用 50 层和 152 层的残差网络，而本文方法中的空间流和时间流均使用 50 层的压缩激励残差网络。对于单个空间流网络来说，时空乘法网络对于 224 像素×224 像素的输入图像单向传播处理需要大约 3.86 GFLO/s（GFLO/s 表示每秒 10 亿次浮点运算）。相较于时空乘法网络，本文方法由于利用了压缩激励操作，故此需要大约 3.87 GFLO/s，增加了大约 0.26%。对于 256 个图像的训练批量，时空乘法网络需要 380 ms，本文方法需要大约 418 ms。虽然本文方法中的空间流网络总参数量较时空乘法网络中的外观流网络需要的 2.5×10^8 的参数量增加了约 10%，但是对于单个时间流网络来说，时空乘法网络运动流使用的 152 层残差网络的网络层数是本文时间流网络层数的 3 倍，而且它的参数量也远多于本文时间流网络的参数量。因此，综合考虑空间流和时间流这两方面，本文方法在总的参数量上少于时空乘法网络，并且在 HMDB51 数据库上本文方法获得了更好的识别效果，在 UCF101 数据库上也达到了和时空乘法网络相媲美的效果。最近提出的语义图像网络方法[29]将扭曲光流和语义光流输入状态细化的长短时记忆网络训练，通过对这 2 种网络的结果求均值得到最终预测结果。本文方法相较于语义图像网络方法在 UCF101 和 HMDB51 数据集上准确率分别提升了 0.3%和 3.5%。综合以上分析可知，本文方法在视频动作识别上具有一定的先进性和优越性。

4　结束语

本文提出了一种时空压缩激励残差乘法网络的动作识别方法。将图像领域的压缩激励网络和残差网络相结合得到的压缩激励残差网络，迁移到时空网络的双流动作识别中。将恒等映射核作为时间滤波器注入到网络模型中，以学习长期时间依赖关系。并对空间压缩激励残差网络和时间压缩激励残差网络进行特征相乘融合，以便更好地学习视频时空特征。此外，通过 3 种不同的策略生成多个模型，并在测试阶段对这些模型结果进行均值以及加权平均法集成以得到最终识别结果。在 HMDB51 和 UCF101 数据集上的识别准确率实验表明，本文方法对动作识别具有良好的性能。本文网络结构采用经典的以 RGB 图像和光流为输入的双流网络结构，下一步的研究工作是探索新的输入方式，以利用多流网络结构进行动作识别。

参考文献：

[1] HERATH S, HARANDI M, PORIKLI F. Going deeper into action recognition: a survey[J]. Image and Vision Computing, 2017(60): 4-21.

[2] 胡琼, 秦磊, 黄庆. 基于视觉的人体动作识别综述[J]. 计算机学报, 2013, 36(12): 2512-2524.
HU Q, QIN L, HUANG Q. Overview of human action recognition based on vision[J]. Chinese Journal of Computers, 2013, 36(12): 2512-2524.

[3] 朱煜, 赵江坤, 王逸宁. 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 Y, ZHAO J K, WANG Y N. A review of human action recognition based on deep learning[J]. ACTA Automatica Sinica, 2016, 42(6): 848-857.

[4] 罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 173-184.
LUO H L, WANG C J, LU F. Survey of video behavior recognition[J]. Journal on Communications, 2018, 39(6): 173-184.

[5] BOBICK A F, DAVIS J W. An appearance-based representation of action[C]//International Conference on Pattern Recognition. IEEE, 1996: 307-312.

[6] WEINLAND D, RONFARD R, BOYER E. Free viewpoint action recognition using motion history volumes[J]. Computer Vision and Image Understanding, 2006, 104(2-3): 249-257.

[7] YILMAZ A, SHAH M. Actions sketch: a novel action representation[C]//Computer Vision and Pattern Recognition. IEEE, 2005: 984-989.

[8] WANG H, ULLAH M M, KLASER A, et al. Evaluation of local spatio-temporal features for action recognition[C]//British Machine Vision Conference. BMVA, 2009: 1-11.

[9] KLASER A, SCHMID C. Action recognition by dense trajectories[C]//Computer Vision and Pattern Recognition. IEEE, 2011: 3169-3176.

[10] WANG H, SCHMID C. Action recognition with improved trajectories[C]//International Conference on Computer Vision. IEEE, 2013: 3551-3558.

[11] 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.

[12] DU T, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//International Conference on Computer Vision. IEEE, 2015: 4489-4497.

[13] TRAN D, RAY J, SHOU Z, et al. ConvNet architecture search for spatiotemporal feature learning[J]. Computing Research Repository, 2017, 16(8): 178-190.

[14] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Computer Vision and Pattern Recognition. IEEE, 2014: 1725-1732.
[15] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Neural Information Processing Systems. NeurIPS, 2014: 568-576.
[16] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[J]. ACM Transactions on Information Systems, 2016, 22(1): 20-36.
[17] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal residual networks for video action recognition[C]//Neural Information Processing Systems. NeurIPS, 2016: 3468-3476.
[18] WANG X, FARHADI A, GUPTA A. Actions~transformations[C]// Computer Vision and Pattern Recognition. IEEE, 2016: 2658-2667.
[19] WANG Y, LONG M, WANG J, et al. Spatiotemporal pyramid network for video action recognition[C]//Computer Vision and Pattern Recognition. IEEE, 2017: 2097-2106.
[20] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//Computer Vision and Pattern Recognition. IEEE, 2016: 1933-1941.
[21] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal multiplier networks for video action recognition[C]// Computer Vision and Pattern Recognition. IEEE, 2017: 7445-7454.
[22] WANG L, GE L, LI R, et al. Three-stream CNNs for action recognition[J]. Pattern Recognition Letters, 2017, 92(C): 33-40.
[23] BILEN H, FERNANDO B, GAVVES E, et al. Action recognition with dynamic image networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(12): 2799-2813.
[24] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Computer Vision and Pattern Recognition. IEEE, 2016: 770-778.
[25] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Computer Vision and Pattern Recognition. IEEE, 2018: 7132-7141.
[26] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. Computer Science, 2012, 3(12): 1-9.
[27] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//International Conference on Computer Vision. IEEE, 2011: 2556-2563.
[28] ZHANG C L, ZHANG H, WEI X S, et al. Deep bimodal regression for apparent personality analysis[C]//European Conference on Computer Vision Workshops. Springer, 2016: 311-324.
[29] KHOWAJA S A, LEE S-L. Semantic image networks for human action recognition[J]. The Computing Research Repository, 2019, 21(1): 1-30.

[作者简介]

罗会兰（1974− ），女，江西上高人，博士，江西理工大学教授，主要研究方向为计算机视觉、模式识别。

童康（1992− ），男，江苏南京人，江西理工大学硕士生，主要研究方向为计算机视觉、视频动作识别。