计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于深度学习的轻量型的人体动作识别模型
作者：何冰倩，魏维，张斌
DOI：10.19734/j.issn.1001-3695.2019.02.0094
收稿日期：2019-02-16
网络首发日期：2019-07-10
引用格式：何冰倩，魏维，张斌．基于深度学习的轻量型的人体动作识别模型[J/OL]．计算机应用研究. https://doi.org/10.19734/j.issn.1001-3695.2019.02.0094

网络首发时间：2019-07-10 10:54:29
网络首发地址：http://kns.cnki.net/kcms/detail/51.1196.TP.20190708.1512.017.html

基于深度学习的轻量型的人体动作识别模型 *

何冰倩[a]，魏 维[b†]，张 斌[a]

(成都信息工程大学 a. 计算机学院; b. 软件工程学院，成都 610225)

摘 要：针对现有基于深度学习的人体动作识别模型参数量大、网络过深过重等问题，提出了一种轻量型的双流融合深度神经网络模型并将该模型应用于人体动作识别。该模型将浅层多尺度网络和深度网络相结合，实现了模型参数量的大幅减少，避免了网络过深的问题。在数据集 UCF101 和 HMDB51 上进行实验，该模型在 ImageNet 预训练模式下分别取得了 94.0%和 69.4%的识别准确率。实验表明，相较于现有大多基于深度学习的人体动作识别模型，该模型大幅减少了参数量，并且仍具有较高的动作识别准确率。

关键词：深度学习；图像处理；卷积神经网络；动作识别

中图分类号：TP391.41 doi: 10.19734/j.issn.1001-3695.2019.02.0094

Lightweight human action recognition model based on deep learning

He Bingqian[a], Wei Wei[b†], Zhang Bin[a]

(*a. School of Computer Science, b. School of Software Engineering, Chengdu University of Information Technology, Chengdu* 610225, *China*)

Abstract: Aiming at the problems that the existing human motion recognition methods based on deep learning have large parameters and the networks were too deep and heavy, this paper proposed a lightweight two-steam fusion deep neural network model and applied this model to human action recognition. This model combined a shallow multi-scale network with a deep network, and achieved a significant reduction in the amount of model parameters and avoided the problem that network was too deep. Experiments were performed on datasets UCF101 and HMDB51, achieving 94.0% and 69.4% recognition accuracy in ImageNet pre-training mode, respectively. Experiments show that compared with the existing human motion recognition models based on deep learning, this model greatly reduces the parameter quantity and still has high motion recognition accuracy.

Key words: deep learning; image processing; convolutional neural network; action recognition

0 引言

人体动作识别长期以来一直是计算机视觉领域的热点研究课题之一[1]，在人机交互[2]、智能家居、医疗复健、智能视频监控等方面也都有着广泛的应用。在这一研究领域，国内外学者提出了许多针对人体动作识别任务的方法和模型，其中基于深度学习的方法在一些公开动作数据集上取得的成果较为显著。但是由于视频中光照、摄像机角度及遮挡、视频背景变化及背景复杂度等因素，基于视频的人体动作识别仍然是一个十分具有挑战性的研究课题。人体动作识别任务一般包含两个阶段：动作特征的提取和表示、动作的建模和识别。动作特征的提取和表示是影响人体动作识别准确率的关键步骤，其方法主要可以分为基于手动提取特征的方法[3~5]和基于深度学习的特征提取方法[6,7]。而根据动作特征的提取和表示的方式不同，选择对动作进行建模和识别的方法也不尽相同，文献[8]将其大致分为三种方法：基于模板的方法、基于概率统计的方法[3, 9]、基于语法的方法[1]。在基于深度学习的人体动作识别方法中大多数模型采用基于概率统计的判别式模型来对人体动作进行分类。

自 Yann 等人[10]提出 LeNet 网络，并在手写数字识别任务上取得可观的成果后，国内外学者相继提出各种基于深度学习的网络模型并应用于人体动作识别，比如 Alex 等人[11]提出的 AlexNet，Simonyan 等人[12]提出的 VggNet，Szegedy 等人[13]提出的 GoogleNet，何凯明等人[14]提出的 ResNet，Huang 等人[15]提出的 DenseNet 等等。AlexNet 和 VggNet 均是通过加深网络深度的方式来提高网络性能，GoogleNet 和 ResNet 采用增加网络模型的宽度或深度的方式来提高网络性能，这些网络在图片识别和分类以及人体动作识别等领域均取得了非常可观的成绩[16]。尽管文献[12]通过对网络模型增加三层权重层的实验证明浅层学习网络对复杂函数的表示以及模型的泛化能力均具有一定的局限性，但是随着网络层数的不断叠加和宽度的不断扩展，也会带来一些问题，例如参数量巨大、网络的计算复杂度较大、网络越深越容易出现梯度消失等等，文献[14]也提到盲目增加网络深度，会出现准确率包含或者下降的情况，从而带来网络模型的退化问题。因此，研究如何利用轻量级的模型对人体动作进行分类而不损失精度逐渐成为许多学者的新的研究方向。文献[17]提出了一种深度网络结构模块（Network In Network, NIN），NIN 模块利用多层感知模型能够增强视野接受域内对局部区域的辨别能力。Liu 等人[18]受人类视觉的感受域的启发，提出了 RFB（Receptive Field Block）模块，该模型是以 Vgg16[12]为主干网络进行构建的轻量型网络模型，RFB 模块主要由多分支卷积层和连接在卷积层后的膨胀卷积层组成，在图片分类领域达到了精度和速度并举的要求，但当该模型应用于基于

收稿日期：2019-02-16；**修回日期：**2019-03-29 **基金项目：**四川省教育厅重点科研项目（17ZA0064）

作者简介：何冰倩(1994-)，女，四川阆中人，硕士研究生，主要研究方向为图形图像处理；魏维(1976-)，男(通信作者)，四川西昌人，教授，硕导，博士，主要研究方向为图形图像处理及应用(weiwei@cuit.edu.cn)；张斌(1992-)，男，四川巴中人，硕士研究生，主要研究方向为图形图像处理.

视频的人体动作识别任务时，因为其缺乏对时间上的运动信息的分析，具有一定的局限性[16]。

基于视频的人体动作识别任务与基于静态图像的图片识别任务的主要区别在于视频序列不仅包含图像的外观信息还包含时间序列上的运动信息，而单图像的分析识别不需要去考虑时间上的信息，因此，为了弥补二维卷积神经网络模型不能有效结合视频序列中的运动信息，文献[19]提出三维卷积神经网络模型并将该模型应用于人体动作识别任务。文献[20]关注于外观信息和运动信息的有效结合，提出了双流卷积神经网络模型（Two-stream Convolutional Networks），之后也有许多在该模型基础上进行改进并取得不错效果的人体动作识别模型[21, 22]，该模型的输入是视频序列的 RGB 数据和光流（Flow）数据。文献[23]提出了更深的卷积神经网络模型（Convolutional 3D），称为 C3D 模型，该模型近似于三维版本的 VggNet 模型。这些模型在一定程度上考虑了视频序列具有运动信息的特性，但是仍然存在关注提高识别准确率的精度时网络结构仍然变得越来越深的问题。

针对上述问题，本文提出了一种浅层和深层网络相结合的轻量型的深度学习网络（lightweight deep learning network model combining shallow and deep networks, SDNet），并将其应用于人体动作识别。为了更好的提取特征和增强模型的泛化能力，本文提出了浅层多尺度模块，并对 NIN 深层网络模块进行改进，然后将浅层网络和深层网络进行结合。同时，由于基于视频的人体动作识别具有一定时间长度上的运动信息的特性，为了使模型更好地利用视频序列的时空信息，本文采用三维卷积操作来实现对时空信息的提取。从而实现了浅层和深层网络相结合的轻量型的深度学习网络的设计和构建。本文根据双流卷积神经网络模型[20]的思想，利用本文提出的 SDNet 构建了基于深度学习的轻量型的人体动作识别模型。

本文在公开数据集 UCF101 和 HMDB51 上进行实验，实验结果表明本文提出的基于深度学习的轻量型的人体动作识别模型能够较好的兼顾精度和速度的要求，且能有效识别视频序列中的人体动作。

1　浅层和深层网络相结合的轻量型的深度学习网络设计

1.1　整体网络设计

本文提出的浅层和深层网络相结合的轻量型的深度学习网络（SDNet）如图 1 所示。本网络模型主要包含两大模块，一个是浅层多尺度网络模块，一个是深层网络模块。具体来说，本文模型由一个卷积核为 3×3×3 的卷积层、一个最大池化层、三个密集连接的浅层多尺度模块、一个连接作用的卷积核为 1×1×1 的卷积层、一个深层网络模块、一个全连接层和 softmax 组成。密集连接的浅层多尺度模块主要负责对视频序列的局部特征进行提取和组合，形成较长较广的深层特征；深层网络模块的主要作用是凭借其较好的抽象能力更好的融合上一模块提取到的特征，增强整个网络模型对特征的表达能力。浅层网络和深层网络的结合，使得本文模型对人体动作识别任务在网络模型不厚重的前提下，能够对视频序列的时空特征更好的表征并取得较好的识别结果。

1.2　浅层多尺度模块

一般而言，动作特征的提取可以按照是否是全局特征然后分为基于全局的特征提取以及基于局部的特征提取。而对于基于视频的人体动作识别任务而言，存在视频背景大多是复杂环境或者部分遮挡的问题，基于全局的特征提取并不能很好的应对这种情况[24]，而多尺度表达可以利用简单的方式对视频序列中的局部特征进行不同尺度的描述，便于分析视频帧的局部特征，因此本文模型采用多尺度特征提取的方式对视频序列进行局部特征提取。

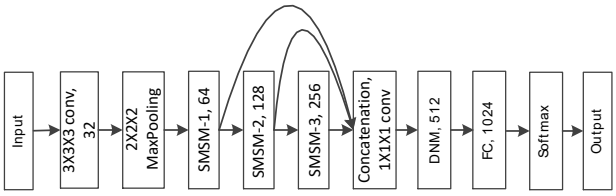

图 1　浅层和深层网络相结合的轻量型的深度学习网络

Fig. 1　Lightweight deep learning network model combining shallow and deep networks, sdnet

本文受 RFB 模块[18]的启发，利用膨胀卷积[25]操作来扩大网络模型的接受域，提出了基于浅层网络的多尺度模块（Shallow multi-scale module, SMSM），该模块如图 2 所示，图 2 是以图 1 模型中 SMSM-1 模块为例。该模块中 1×1×1 的卷积层不包含非线性激活函数，其余卷积层均采用 ReLU 作为非线性激活函数。具体来说，本文首先利用 1×1×1 的卷积层组成模块中每个分支的瓶颈结构（bottleneck structure），采用该结构的主要目的是减少对下一层输入的特征映射的通道数目，从而提高计算效率。然后利用两个叠加的 3×3×3 卷积层替代 5×5×5 的卷积层，该操作可以有效减少模型参数。最后将不同膨胀系数的卷积层提取到的特征连接起来，通过最大池化层连接到模型的下一结构。

基于视频的人体动作识别往往包含较长期的时空依赖关系，而膨胀卷积操作对这类问题有较好的应用[25]，同时该操作可以在不做池化操作而损失信息并且保持相同参数的情况下，通过加大感受野，使得卷积操作能够输出较大范围的上下文特征信息，同时，经过文献[25]的证明该操作可以提高计算速度和识别准确率。本文对 SMSM 模块的每个分支中两个叠加的 3×3×3 卷积层均进行膨胀操作，如图 2 所示，每个分支的膨胀系数分别设置为 $\{1,2,3\}$ 。同时鉴于文献[18]中膨胀系数分别设置为 $\{1,3,3\}$ ，本文在实验阶段设置了两组不同膨胀系数的对比实验。

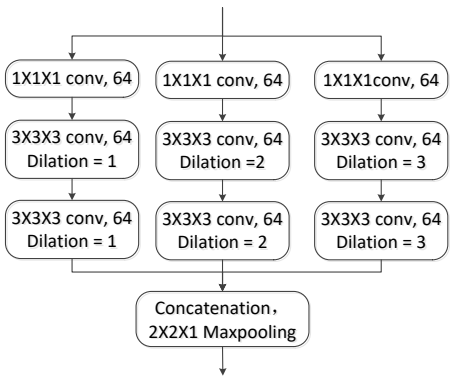

图 2　基于浅层网络的多尺度模块（SMSM）

Fig. 2　Shallow multi-scale module, SMSM

1.3　深层网络模块

文献[15]经过实验证明，密集连接的网络不仅不会带来冗余的问题，还可以减轻梯度消失的问题，使网络的泛化能力得到增强，其原因主要是密集连接极大地减少了每一层的计算量，并且除了密集连接的最后一层外其余每一层提取到的特征都能被重复利用。由此可见密集连接能够提高网络模型的性能，因此本文将三个不同滤波器数目的浅层多尺度模块的特征进行密集连接，从而形成较长较广的深层特征。为

了更好的融合浅层多尺度模块（SMSM）提取到的多尺度特征，本文在深度网络结构模块[17]（network in network, NIN）的基础上进行改进，并利用改进后的 NIN 模块形成本文模型中的深层网络模块。

NIN 模块具有较高的抽象水平，即就人体动作识别任务而言，NIN 模块可以在整个模型的低层特征形成高层特征的时候，能尽可能得到同一动作在不同角度和尺度下仍保持不变的特征，这一较高的抽象水平能力可以增强本文人体动作识别模型对局部特征的表达能力。文献[17]中 NIN 网络如图 3 所示，该网络可以近似的看做是由叠加的三个 NIN 模块和一个全局平均池化层组成的。每一个 NIN 模块包含一个 3×3 的卷积层、一个 1×1 的卷积层和多层感知机的组合（MLPConv）、一个 3×3 的卷积层。其中，MLPConv 和其余卷积层一样均以 Relu 作为其激活函数。MLPConv 较之简单线性卷积层具有更强的对各种潜在概念分布建模的能力，能够提升传统 CNN 网络的特征表达能力。

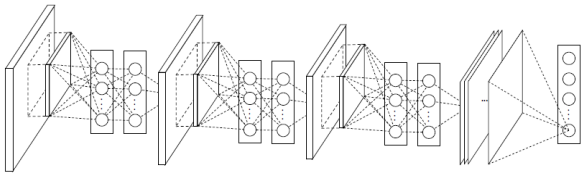

图 3　NIN 网络模型

Fig. 3　The overall structure of Network In Network

本文的深层网络模块如图 4 所示。NIN 网络本身是由多个 Conv+MLPConv+Conv 这样的结构叠加而成，而本文采用两个这样的结构来叠加形成本文的深层网络模块。具体来说，本文首先对原 NIN 网络进行了两方面的改进：一方面是考虑到本文的人体动作识别任务中尽可能捕获充分时空信息的要求，将原 NIN 的各层卷积和池化均扩展为三维操作；另一方面是由于对 SDNet 网络上一结构浅层多尺度模块大范围使用了扩大感受野操作，因此考虑到浅层多尺度模块密集连接后的多尺度特征的有效融合，对深层网络模块的第一个 Conv+MLPConv+Conv 结构做进一步改进，即对该第一个结构的所有卷积操作添加膨胀系数为 2 的膨胀卷积操作。然后通过 2×2×2 的最大池化层，将第一个改进后的结构连接到第二个改进后的结构，其中第二个改进后的结构中所有卷积的卷积核膨胀系数为 1。最后考虑到全局平均池化层能够减少参数量，而且能较好地弥补全连接层容易过拟合的缺点，符合本文试图构建轻量型的人体动作识别模型的要求，因此仍采用全局平均池化在分类层作为特征映射的池化方式。

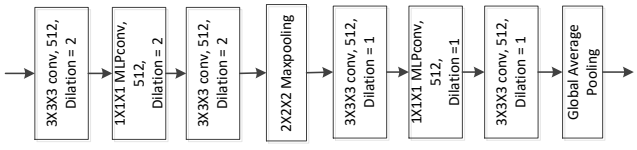

图 4　SDNet 的深层网络模块（DNM）

Fig. 4　Deep networks module of sdnet, DNM

2　基于深度学习的轻量型的人体动作识别模型设计

2.1　整体框架构建

本文将提出的浅层和深层网络相结合的轻量型的深度学习网络应用到人体动作识别任务，考虑到包含人体动作的视频序列不仅包含空间序列上的外观信息，也包含时间序列上的运动信息，因此采用双流网络模型[20]的概念来构建本文的基于深度学习的轻量型的人体动作识别模型（lightweight human motion recognition model based on deep learning，DLLM）。

文献[26]中表示较长的三维卷积会通过降低输入视频序列的分辨率来减少内存消耗，但是这种操作会丢失一部分的时空线索。为了尽可能不丢失时空线索，从而充分利用时空信息，同时为了将 SDNet 更好的融合到本文的人体动作识别模型中，本文采用时间金字塔池化层（temporal pyramid pooling，TPP）作为 SDNet 和双流网络的全连接层的衔接。时间金字塔池化层能够将帧级特征编码为具有多时间尺度的固定尺寸的视频级表示，采用从粗到细的结构捕捉视频中的人体动作的时间结构，从而更好地衔接 SDNet，增强人体动作识别模型的识别能力和泛化能力。

本文构建的人体动作识别模型 DLLM 如图 5 所示。首先，本文利用之前提出的 SDNet 对时空双流进行特征提取和表示，此处的 SDNet 不包含原网络中的最后一个全连接层和 softmax 层。然后利用时间金字塔池化层（TPP）将时间流和空间流的视频帧级的特征聚合成视频级表示，再通过全连接层和 softmax 层得到时空双流对输入序列的识别结果，最后，利用加权平均融合的方式对双流结果进行融合，从而得到最终的识别结果。

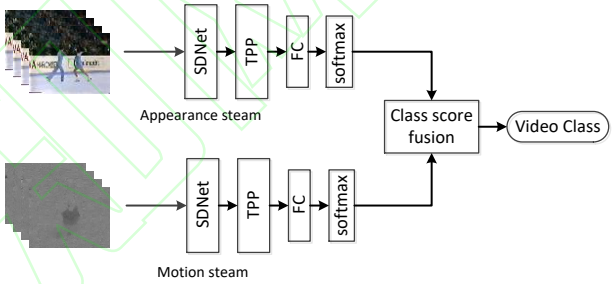

图 5　基于深度学习的轻量型的人体动作识别模型

Fig. 5　Lightweight human motion recognition model based on deep learning, DLLM

2.2　模型预训练

对于 DLLM 模型的预训练，本文考虑两种方式：一种是利用常应用于大多数目标识别和检测的 ImageNet[27]图像数据集对模型进行预训练；另一种方式是采用 Kinetics[28]数据集作为模型预训练的数据集。

ImageNet 预训练。对于模型的空间流，由于其输入为 RGB 数据，因此直接使用 ImageNet 数据集对空间流进行预训练。对于模型的时间流，由于其输入为光流数据，本文利用文献[29]中时间流的预训练方式，即将视频序列的光流场通过线性变换离散到[0~255]内，以保证和 RGB 数据同区间，再根据时间流的输入通道数量对 RGB 通道上权重进行平均后的平均值进行复制，作为时间流的初始化数值。

Kinetics 预训练。Kinetics 数据集是一个相较于 UCF101[30]和 HMDB51[31]而言具有较大规模的视频数据集，该数据集包含了 600 类的人体动作，每个类别包含至少 600 段视频。文献[16]利用 Kinetics 数据集对大多数经典模型框架，例如双流模型、C3D 模型等，重新进行了评估，实验证明通过大规模数据集预训练后的模型框架对较小基准的视频数据集在性能上有明显的提高。因此，本文以 Kinetics 视频数据集作为本文模型预训练的第二种方式。

3　实验与分析

3.1　实验数据集及评估指标

本文实验采用的数据集为国际公开的两个基于视频的行为识别数据集：UCF101 和 HMDB51。UCF101 数据集是由

美国 University of central Florida（UCF）发布的数据库,该数据集总共包含 13320 个视频段，101 类动作，这 101 个类别的动作视频可以被分成 25 组，每组由 4~7 个动作视频组成，而且来自同一组的视频大多会有一些共同的特点，比如相似的背景，相似的观点等。UCF101 数据集大多是真实情景拍摄，因此存在背景复杂、光照变化等挑战；同时，该数据集涵盖了人与人的交互、人与物的交互、单人或多人运动等较大范围的人体动作,因此还具有较多的类间和类内差异问题，这也是动作识别的一大挑战。HMDB51 数据集是 Brown university 在 2011 年发布的行为识别数据集，该数据集总共包含 6849 段视频样本，分为 51 类动作，每类至少包含有 101 个视频段，其视频段的主要来源是电影剪辑片段，因此对于动作识别任务，存在背景复杂、视频像素低等挑战。在实验阶段，本文按照文献[30,31]中规定的分割集划分标准，将 UCF101 和 HMDB51 这两个数据集分别划分为三个分割子集（split），而每个分割子集大约可以分为 70%的训练集 30%的测试集。

本文采用实验数据集的最终动作识别准确率作为本文人体动作识别模型的评估指标。而最终识别准确率分别是两个实验数据集的三个分割子集的识别准确率经过线性加权平均后得到的值。

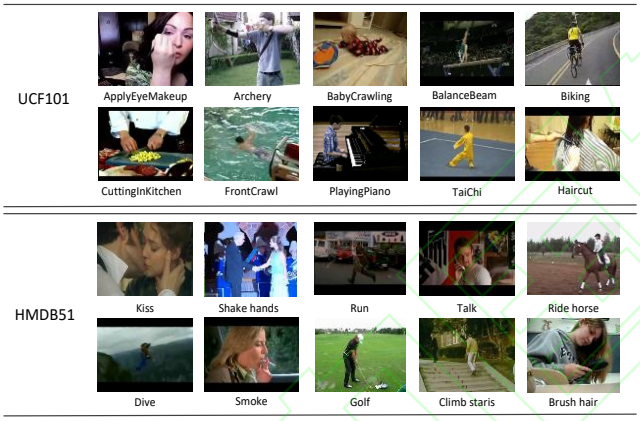

图 6　UCF101 和 HMDB51 数据集的部分动作示例图像

Fig. 6　Sample action images of datasets UCF 101 and HMDB51

3.2　实验结果与分析

本文将人体动作识别模型在 Linux 系统（Ubuntu16.04）单 GPU 环境下搭建的 TensorFlow 平台上进行实验。具体细节及参数如下。本文在视频训练时使用 SGD（stochastic gradient descent）优化算法，动量值设置为 0.9。在时间流和空间流网络中通过 dropout 层尽量避免模型过拟合，dropout 率分别设置为 0.7 和 0.8。空间流的学习率（learning rate）初始化设置为 10^{-2}，当验证损失（validation loss）饱和时，将学习率调整为原来的 0.1 倍，当学习率降为 10^{-5} 时结束训练；时间流的学习率初始化设置为 10^{-3}，与空间流迭代规则一致，当学习率降为 10^{-5} 时结束训练。对于从视频中采样的所有帧序列都采用相同的数据增强方式[29]，包括随机裁剪、水平翻转和比例抖动。时间金字塔池化水平设置为 $\{4\times4\times1,2\times2\times1,1\times1\times1\}$,即时间金字塔采用 3 层金字塔形式。

本文通过 DLLM 模型中的 SDNet 提取视频序列中的时空特征，而在 SDNet 的 SMSM 模块设计阶段，考虑到不同膨胀系数的设置会影响感受域大小，从而影响动作识别准确率，于是本文针对 SMSM 模块的膨胀系数设置设计了对比实验。首先，本文对 SMSM 模块的三个分支考虑两组膨胀系数设置，分别是 {1,2,3} 和 {1,3,5}，然后在数据集 UCF101 的第一分割集（split1）上进行人体动作识别实验，此实验模型的预训练方式为 ImageNet 预训练。实验结果如表 1 所示，在 SMSM 模块中，将三个分支的膨胀系数分别设置为 1、2、3 时，不论是时间流还是空间流的动作识别准确率均优于设置为 {1,3,5} 时的识别准确率。因此，在后续的实验中，每个 SMSM 模块的三个分支的卷积膨胀系数均分别设置为 1、2、3。

表 1　SMSM 模块不同膨胀系数设置下的动作识别准确率对比

Table 1　Action recognition accuracy comparison of different dilation rate settings of SMSM

膨胀系数组	空间流	时间流
{1,2,3}	83.37%	86.06%
{1,3,5}	82.81%	84.50%

表 2 展示了本文的时间流网络和空间流网络在 UCF101 和 HMDB51 上的动作识别准确率。该结果的实验的预训练方式采用的是 ImageNet 预训练方式。根据表 2 可以看出，对于单流的识别准确率，时间域的识别准确率均略高于空间域的。同时这一结果可以表明，就基于视频的人体动作识别任务而言，包含运动信息的时间流比包含外观信息的空间流更为重要，也更具表现力。

表 2　本文时空网络的动作识别准确率

Table 2　Action recognition accuracy of proposed temporal spatial network

分割方式	域	UCF101	HMDB51
Split1	空间域	83.81%	57.68%
	时间域	87.93%	63.62%
Split2	空间域	84.37%	59.06%
	时间域	87.50%	65.18%
Split3	空间域	84.25%	59.37%
	时间域	86.96%	62.50%

表 3　本文模型的人体动作识别准确率

Table 3　Action recognition accuracy of proposed method

分割方式	UCF101	HMDB51
Split1	94.25%	70.31%
Split2	93.87%	68.18%
Split3	93.75%	69.62%
线性平均	93.96%	69.37%

本文的 DLLM 模型在数据集 UCF101 和 HMDB51 上的人体动作识别准确率如表 3 所示。模型的预训练方式为 ImageNet 预训练。每一分割子集（split）的识别准确率均是对模型的空间流和时间流的结果进行决策融合得到的，决策融合时的时间流和空间流的识别置信度为 1:1。然后再对数据集的每一个分割子集进行线性加权平均，得到人体动作识别模型在数据集上的最终识别准确率。对比表 2 可以看出，融合后的时空双流的动作识别准确率要高于单流的时间流或空间流，这也说明将时间流和空间流进行融合才能更好地利用视频序列中的时空信息，从而使得动作识别模型更具有表现力。

表 4　不同模型的网络参数量及时间轴输入大小对比

Table 4　Comparison of number of parameters and temporal input sizes of different models

Model	Params	Training Input Frames	Testing Input Frames
文献[20]	12M	1 RGB, 10 Flow	25 RGB, 250 Flow
文献[23]	65M	16 RGB	240 RGB
文献[32]	32M	5 RGB, 50 Flow	25 RGB, 250 Flow
文献[16]	25M	64 RGB, 64 Flow	250 RGB, 250 Flow
本文方法	19M	64 RGB, 64 Flow	250 RGB, 250 Flow

表 5　不同模型的人体动作识别准确率和参数量对比

Table 5　Action recognition accuracy and parameters comparison of proposed method and other methods

Pre-training	Model	Year	Params	UCF101	HMDB51
ImageNet pre-training	文献[20]	2014	12M	88.0%	59.4%
	文献[23]	2015	65M	85.2%	-
	文献[32]	2016	32M	92.5%	65.4%
	文献[16]	2017	25M	93.4%	66.4%
	本文方法		**19M**	**94.0%**	**69.4%**
Kinetics pre-training	文献[20]	2014	12M	93.4%	63.6%
	文献[23]	2015	65M	91.1%	-
	文献[32]	2016	32M	95.3%	73.5%
	文献[16]	2017	25M	**97.9%**	**80.2%**
	本文方法		**19M**	96.5%	74.6%

本文的主要思想是在尽量不损失精度而且尽可能减少模型参数和计算量的情况下，构建轻量型的精度和轻度并举的人体动作识别模型。因此，本文对近年的经典人体动作识别模型进行了复现和实验，与本文方法做参数量的定量比较如表 4 所示，其中对于输入帧，本文通过保持每 5 帧中有 1 帧来自每秒 25 帧的原始流来进行子采样。这些模型与本文的识别准确率的对比实验结果如表 5 所示。在实验阶段，对每个对比模型均采用了两种预训练方式。Two-Stream 模型是文献[20]在 2014 年提出的，双流网络均是 2D 网络，因此该模型的参数量较少，约为 12M；C3D 模型是文献[23]在 2015 年提出的，该模型是单流的 3D 模型，具有较深的网络结构；Two-steam VGG 模型是文献[32]在 2016 年提出的，该模型双流的 VGG 模型；Two-Stream I3D 模型是文献[16]在 2017 年提出的双流 I3D（Two-Stream Inflated 3D ConvNet）融合模型。从表 4 的参数量对比和表 5 的实验结果可以直观的看出，本文在保持精度水平和目前前沿方法的动作识别准确率大致一致的前提下，模型的参数量明显少于其他方法的参数量。这一结果表明，本文提出的基于深度学习的轻量型的人体动作识别模型能够有效地实现了减少了模型的参数量而不损失精度。

4　结束语

在计算机视觉领域，基于深度学习的方法已经得到了广泛的研究和应用。针对计算机视觉领域中的基于视频的人体动作识别任务，本文提出了浅层和深层网络相结合的轻量型的深度学习网络（SDNet），并以此为网络基础，建立了基于深度学习的轻量型的双流融合人体动作识别模型（DLLM）。本文模型分别在 ImageNet 和 Kinetics 数据集上进行了预训练和参数微调， 然后在公开数据集 UCF101 和 HMDB51 上进行实验，对于两种预训练方式，在数据集 UCF101 上分别取得了 94.0%和 96.5%的动作识别准确率；在数据集 HMDB51 上分别取得了 69.4%和 74.6%的动作识别准确率，而模型参数量仅为 19M。实验结果表明，本文提出的基于深度学习的轻量型的人体动作识别模型不仅能够对视频中的人体动作进行有效识别，还相较于近年人体动作识别模型大幅减少了参数量，节省了计算成本。但是，在本文模型中，由于浅层网络和深层网络的叠加使用，模型的收敛速度较慢。因此，今后可以针对这一问题进行研究改进，提出更具表现力和泛化能力强的轻量型的深度学习网络模型。

参考文献：

[1] Lei Zhang, Zhen Xiantong, Shao Ling, *et al.* Learning match kernels on Grassmann manifolds for action recognition [J]. IEEE Trans on Image Processing, 2019, 28 (1): 205-215.

[2] Piyathilaka L, Kodagoda S. Gaussian mixture mased HMM for human daily activity recognition using 3D skeleton features [C]//Proc of the 8th IEEE Conference on Industrial Electronics and Applications. Piscataway,NJ: IEEE Press, 2013: 567-572.

[3] Najar F, Bourouis S, Bouguila N, *et al.* A fixed-point estimation algorithm for learning the multivariate GGMM: Application to human action recognition [C]// Proc of IEEE Canadian Conference on Electrical & Computer Engineering. Piscataway,NJ: IEEE Press, 2018: 1-4.

[4] 冯小明，冯乃光，汪云云. 基于运动特征与序列袋的人体动作识别 [J]. 计算机工程与设计, 2018, 39 (10): 3220-3227. (Feng Xiaoming, Feng Naiguang, Wang Yunyun. Body motion recognition based on moving feature coupled bag-of-sequence [J]. Computer Engineering and Design, 2018, 39 (10): 3220-3227)

[5] Liu Fang, Xu Xiangmin, Qiu Shuoyang, *et al.* Simple to complex transfer learning for action recognition [J]. IEEE Trans on Image Processing, 2016, 25 (2): 949-960.

[6] Yu Gang, Li Ting. Recognition of Human Continuous Action with 3D CNN [C]//Proc of International Conference on Computer Vision Systems. 2017: 314-322.

[7] 韩敏捷. 基于深度学习框架的多模态动作识别 [J]. 计算机与现代化, 2017 (7): 48-52. (Han Minjie. Multi-modal action recognition based on deep learning framework [J]. Computer and Modernization, 2017 (7): 48-52)

[8] 胡琼，秦磊，黄庆明. 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 qiong, Qin Lei, Huang Qingming. A survey on visual human action recognition [J]. Chinese Journal of Computers, 2013,36 (12): 2512-2524)

[9] Jagadeesh B, Patil C M. Video based action detection and recognition human using optical flow and SVM classifier [C]//Proc of IEEE International Conference on Recent Trends in Electronics: Information & Communication Technology. 2017, 2017: 1761-1765.

[10] LeCun Y, Bottou L, Bengio Y, *et al.* Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86 (11): 2278-2324.

[11] Alex K, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks [C]// Proc of the 25th International Conference on Neural Information Processing Systems. 2012: 1097-1105

[12] Karen S, Andrew Z. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014, arXiv: 1409. 1556v6.

[13] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2015: 1-9.

[14] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep Residual Learning for Image Recognition [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2016: 770-778.

[15] Gao Huang, Zhuang Liu, Laurens van der Maaten, *et al.* Densely connected convolutional networks [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 2261-2269.

[16] João Carreira, Andrew Zisserman. Quo Vadis, Action recognition?A

new model and the kinetics dataset [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2017: 4724-4733.

[17] Lin Min, Chen Qiang, Yan Shuicheng. Network in network [J]. arXiv. org, 2013: arXiv: 1312. 4400 [cs. NE].

[18] Liu Songtao, Huang Di, Wang Yunhong. Receptive field block net for accurate and fast object detection [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2018, 11215: 404-419.

[19] Ji Shuiwang, Yang Ming, Yu Kai. 3D convolutional neural networks for human action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (1): 221-231.

[20] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 1 (4): 568-576.

[21] Gammulle Harshala, Denman Simon, Sridharan Sridha, *et al.* Two-stream LSTM: A deep fusion framework for human action recognition [C]//Proc of IEEE Winter Conference on Applications of Computer Vision.Piscataway,NJ: IEEE Press, 2017: 177-186.

[22] Khong V M, Tran T H. Improving human action recognition with two-stream 3D convolutional neural network [C]//Proc of the 1st International Conference on Multimedia Analysis and Pattern Recognition. 2018: 1-6.

[23] Du Tran, Lubomir Bourdev, Rob Fergus, *et al.* Learning spatiotemporal features with 3D convolutional networks [C]//Proc of IEEE International Conference on Computer Vision. Washington DC: IEEE Computer Society, 2015: 4489-4497.

[24] 王刘涛, 王建玺, 鲁书喜. 基于 AdaBoost 关键帧选择的多尺度人体动作识别方法 [J]. 重庆邮电大学学报: 自然科学版, 2015, 27 (4): 549-555. (Wang Songtao, Wang Jianxi, Lu Shuxi. Multi-scale human action recognition method based on AdaBoost key-frame selecting [J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition, 2015, 27 (4): 549-555.)

[25] Yu F, Koltun Vladlen. Multi-scale context aggregation by dilated convolutions [C]//Proc of International Conference on Learning Representations. Caribe Hilton, San Juan, Puerto Rico, 2016: 13.

[26] Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2017, 40 (6): 1510 – 1517.

[27] Deng Jia, Dong Wei, Socher R, *et al.* ImageNet: A large-scale hierarchical image database [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 248-255.

[28] Kay W, Carreira J, Simonyan K, *et al.* The kinetics human action video dataset [J]. arXiv. org, 2017: arXiv: 1705. 06950v1 [cs. CV].

[29] Wang Limin, Xiong Yuanjun, Wang Zhe, *et al.* Temporal segment networks: Towards good practices for deep action recognition [C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 20-36.

[30] Soomro K, Zamir A R, Shah M. UCF101: a dataset of 101 human action classes from videos in the wild [J]. arXiv. org, 2012: arXiv: 1212. 0402 [cs. CV].

[31] Hilde K, Hueihan J, Rainer S, *et al.* HMDB51: a large video database for human motion recognition [C]// High Performance Computing in Science and Engineering. Berlin: Springer, 2013: 571-582.

[32] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway. NJ: IEEE Press, 2016: 1933-1941.