Opto-Electronic Engineering
光 电 工 程

Article

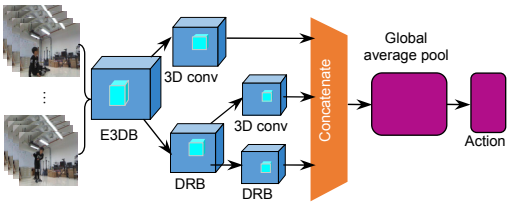
2020 年，第 47 卷，第 2 期

DOI: 10.12086/oee.2020.190139

高效 3D 密集残差网络及其在人体行为识别中的应用

李梁华，王永雄*

上海理工大学光电信息与计算机工程学院，上海 200093



摘要：针对 3D-CNN 能够较好地提取视频中时空特征但对计算量和内存要求很高的问题，本文设计了高效 3D 卷积块替换原来计算量大的 3×3×3 卷积层，进而提出了一种融合 3D 卷积块的密集残差网络(3D-EDRNs)用于人体行为识别。高效 3D 卷积块由获取视频空间特征的 1×3×3 卷积层和获取视频时间特征的 3×1×1 卷积层组合而成。将高效 3D 卷积块组合在密集残差网络的多个位置中，不但利用了残差块易于优化和密集连接网络特征复用等优点，而且能够缩短训练时间，提高网络的时空特征提取效率和性能。在经典数据集 UCF101、HMDB51 和动态多视角复杂 3D 人体行为数据库(DMV action3D)上验证了结合 3D 卷积块的 3D-EDRNs 能够显著降低模型复杂度，有效提高网络的分类性能，同时具有计算资源需求少、参数量小和训练时间短等优点。

关键词：机器视觉；卷积神经网络；行为识别；视频分类

中图分类号：TP391.4　　**文献标志码：**A

引用格式：李梁华，王永雄. 高效 3D 密集残差网络及其在人体行为识别中的应用[J]. 光电工程，2020，**47**(2): 190139

Efficient 3D dense residual network and its application in human action recognition

Li Lianghua, Wang Yongxiong*

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Abstract: In view of the problem that 3D-CNN can better extract the spatio-temporal features in video, but it requires a high amount of computation and memory, this paper designs an efficient 3D convolutional block to replace the 3×3×3 convolutional layer with a high amount of computation, and then proposes a 3D-efficient dense residual networks (3D-EDRNs) integrating 3D convolutional blocks for human action recognition. The efficient 3D convolutional block is composed of 1×3×3 convolutional layers for obtaining spatial features of video and 3×1×1 convolutional layers for obtaining temporal features of video. Efficient 3D convolutional blocks are combined in multiple locations of dense residual network, which not only takes advantage of the advantages of easy optimization of residual blocks and feature reuse of dense connected network, but also can shorten the training time and improve the efficiency and performance of spatial-temporal feature extraction of the network. In the classical data set UCF101, HMDB51 and the dynamic multi-view complicated 3D database of human activity (DMV action3D), it is verified that the 3D-EDRNs combined with 3D convolutional block can significantly reduce the complexity of the model, effec-

收稿日期：2019-03-27； 收到修改稿日期：2019-06-23

基金项目：国家自然科学基金资助项目(61673276，61603255，61703277)

作者简介：李梁华(1994-)，男，硕士研究生，主要从事计算机视觉的研究。E-mail：1244094457@qq.com

通信作者：王永雄(1970-)，男，博士，教授，主要从事智能机器人及视觉的研究。E-mail：wyxiong@usst.edu.cn

tively improve the classification performance of the network, and have the advantages of less computational resource demand, small number of parameters and short training time.

Keywords: machine vision; convolutional neural network; action recognition; video classification

Citation: Li L H, Wang Y X. Efficient 3D dense residual network and its application in human action recognition[J]. *Opto-Electronic Engineering*, 2020, **47**(2): 190139

1 引 言

当今网络大量传播各种文本、图像和视频，特别是随着移动设备的大量普及，图像和视频成为人与人之间一种新的交流通信方式，因此基于多媒体自动理解的 AI 技术不断发展并开始大量使用。近年来卷积神经网络(convolutional neural network，CNN)飞速兴起，其网络的深度和结构越来越多样化。尤其是在图像识别中，表现出了惊人的学习能力，逐步超过人类的识别能力。例如，在 Imagenet 测试集上残差网络实现了 3.57%的错误率，已经低于目前人类 5.1%的错误率[1]。视频序列中不但包括空间特征，而且还有时序特征，对视频分类效果的好坏很大程度上取决于能否从视频中提取和利用这两个特征。有效地从人体行为视频中提取到具有高区分度的时空特征，对于提高人体行为识别的准确率有着重要作用。然而，视频是大量的连续帧序列，具有极大的变化性和复杂性。

为了获得视频中运动信息的时空特征，大量的方法被提出来。例如，HOG3D(histogram of gradient 3D)[2]，SIFT3D(scale invariant feature transform 3D)[3]，HOF(histograms of oriented optical flow)[4]，ESURF (efficient speed up robust features)[5]，IDTs (improved dense trajectories)[6]等，以上方法都是人工设计获取特征。其中，IDTs 取得了很好的效果，但需要消耗大量的计算资源，并且缺少捕获语义概念的功能。当对视频中的时空信息进行编码时，自然的想法是将卷积神经网络的卷积核从 2D 卷积扩展为 3D 卷积。最近，研究人员提出了几种用于视频分类的 3D 时域结构网络，Simonyan 等[7]提出了一种双流网络，即将空间流和时间流结合在一起的网络。Tran 等[8]探讨出 3D 卷积滤波器的大小为 3×3×3。最常见的是 3D 卷积($s\times s\times d$)，其中 d 是卷积核的时间深度，即一次输入的帧数，s 是卷积核空间的大小。虽然使用 3D 卷积可以同时获取时间和空间两个维度的特征，但计算的成本和需要的计算机内存都很大。另一种解决方案是利用池化策略或递归神经网络(RNN)表示视频时空特征[9]，通过激活 2D 卷积神经网络的最后一层池化层或全连接层。但是这种方法只是对网络的高层特征进行时序特征提取，对于浅层的时序特征并未充分利用。

残差网络(ResNet)在分类、定位、检测等方面取得了很好的效果[10]，网络可以扩展到数千个层，并且仍然具有良好的性能。另外，残差网络还可以通过使用批量标准化(batch normalization，BN)[11]来减少梯度消失对网络的影响，降低网络训练过程中的退化程度。但是深度残差网络因网络简单堆叠残差块，而存在训练速度慢的问题。为了进一步加强对卷积层间信息流的利用，Huang 等[12]引入了密集连接的卷积网络(DenseNet)，该网络中的每一层都直接连接到后续的所有层，可以将扩展的重复特性应用到整个网络当中，卷积层间的信息流也可以顺利传输到每一层。深层次的 DenseNet 存在大量连接，不可避免地消耗了大量 GPU 内存。Song 等[13]将卷积层的滤波器数目减半，并将残差块中的两个卷积求和作为输入，得到了密集残差网络(DRNs)。

综上，为弥补 3D 卷积对计算内存需求大等缺陷，本文利用时空卷积滤波器构建了一种特征提取块，称为高效 3D 卷积块(efficient 3D convolution block, E3DB)。根据双流网络两个分支同时获得时空特征的优点，E3DB 将计算量大的 3D 卷积滤波器(大小为 3×3×3)设计为 2D CNN 在空间域作用的 1×3×3 卷积滤波器和 1D CNN 在时间域作用的 3×1×1 卷积滤波器。针对深度神经网络存在训练速度慢、无法有效提取视频中具有多变和复杂的时空特征等问题，文中提出了一种融合 3D 卷积块的密集残差网络(3D-efficient dense residual networks，3D-EDRNs)，该网络结合残差网络易于优化和密集网络特征复用的特性，为视频行为识别提供了一种高效的网络框架，具有计算速度快、参数量小和分类性能好等优点。

主要贡献如下：

1) 提出了 E3DB，E3DB 作为一种新颖的卷积块可以替代标准的 3×3×3 卷积层，能够大幅减少网络的运算量，降低计算资源需求。将 E3DB 应用于经典 C3D

模型，可以将模型的参数量降低一倍，网络的分类性能得到提升；

2) 提出的新型 3D-EDRNs 网络充分利用密集网络易于优化和残差网络特征复用等特性，把 3D-EDRNs 提取的特征输入线性支持向量机(linear SVM)进行人体行为识别，达到的准确率为 97.09%，相比于未加入 E3DB 的 3D-EDRNs 网络提升了 8.79%，准确率比经典 C3D 网络高 14.79%。3D-EDRNs 整个模型的参数量只有 3.43 M，仅为 P3D 网络参数量的 1/28，极大地降低了对计算机内存的使用率。3D-EDRNs 网络预测一个视频段(16 帧)消耗时间为 11 ms，比 C3D 网络运行速度快一倍，在同等计算量的情况下，该网络可以显著降低对计算资源的需求，更快实现视频分类功能。

2 3D 密集残差网络

2.1 C3D 网络

针对 2D 卷积不能很好地捕获视频的时序信息，文献[15]提出了将一段视频序列进行卷积的 3D 卷积核，即将多个连续帧堆叠成一个立方体，在立方体上运用 3D 卷积核。C3D[16]网络作为一个经典的 3D 网络，在行为识别、场景识别、视频相似度分析等领域取得很好的效果。

本文采用了 Caffe 实现的 C3D 最新结构，如图 1 所示。该模型共有 5 个 3D 卷积层，卷积核的数量依次为 64、128、128、256、256，每一个 3D 卷积层后接一个 3D 最大池化层，除第一个池化层的大小为(2，2，1)以外，其余池化层的大小均为(2，2，2)，最后是三个全连接层，前两个全连接层的神经元个数均为 2048，最后一层为 101(视频类别输出数量)。

2.2 密集残差结构

2.2.1 理论基础—残差网络和密集网络

残差网络解决了梯度消失问题，可扩展到数千个层，并且能够保持网络的良好性能。残差网络由众多堆叠的残差块组合而成，每一个残差块可以表示为

$$x_{l+1}=x_l+F(x_l,W)\quad , \tag{1}$$

其中：x_l 和 x_{l+1} 分别为第 l 个残差块的输入和输出，F 是残差函数，W 是第 l 个残差块的参数，图 2(a)为包含快捷路径的原始残差块。残差网络的连接方式可以保证信息流从浅层顺利地传输到较深层，从而有效地降低网络的训练难度，提高网络性能。

为了进一步的改善网络层之间的信息流，密集网络[17]提出了一种新的连接方式，将第 l 层之前所有层都作为第 l 层的输入，图 2(b)为包含密集块的网络。由于密集网络的参数可以得到更充分地利用，其性能明显优于其他参数量相近的模型。密集块可以表示为

$$x_l=H_l([x_0,x_1,...,x_{l-1}])\quad , \tag{2}$$

其中：$[x_0,x_1,...,x_{l-1}]$ 是第 l 层之前所有层输出张量的集合，H_l 是关于密集连接块的复合函数。通过这种连接方式，即使是网络的最后一层也可以直接获得网络第一层的输入信息，网络的训练难度得到降低，过拟合现象也得到有效遏制。

2.2.2 提出的密集残差结构

本文结合残差块易于优化和密集连接特征高效利用的优点，提出了一种新的小型密集残差结构，如图 2(c)所示。新型的密集残差结构将原有密集残差结构从 2D 拓展为 3D，并且融入了 E3DB，可以加速网络训练和提高残差网络的性能，该结构可以表示为

$$x_{l+1}=x_l+F(H_d([x_0,...,x_d]),[W_0,...,W_d])\quad , \tag{3}$$

其中：x_l 和 x_{l+1} 分别为网络中第 l 个密集连接残差块的输入和输出，$[x_0,...,x_d]$ 是第 l 块中从第 0 个到第 d 个卷积层特征映射的连接，F 是残差函数，$[W_0,...,W_d]$ 是第 l 块中所有的参数。

从密集残差的结构图来看，它与残差网络非常相似，他们的区别仅仅在于残差块求和层之前的输入，却导致两种网络架构的性能显著不同。密集残差结构将多变和复杂的浅层特征输入到求和层，为网络提供

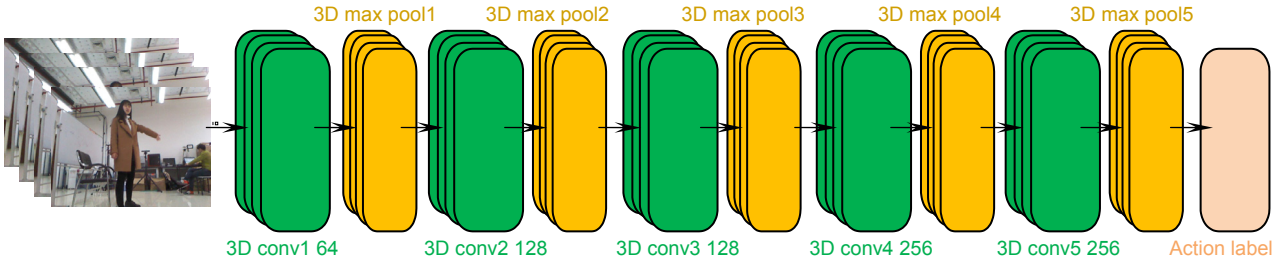


图 1 C3D 网络架构

Fig. 1 C3D network architecture

了更加有效的时空信息，网络的参数更易优化，有利于提升网络的分类性能。

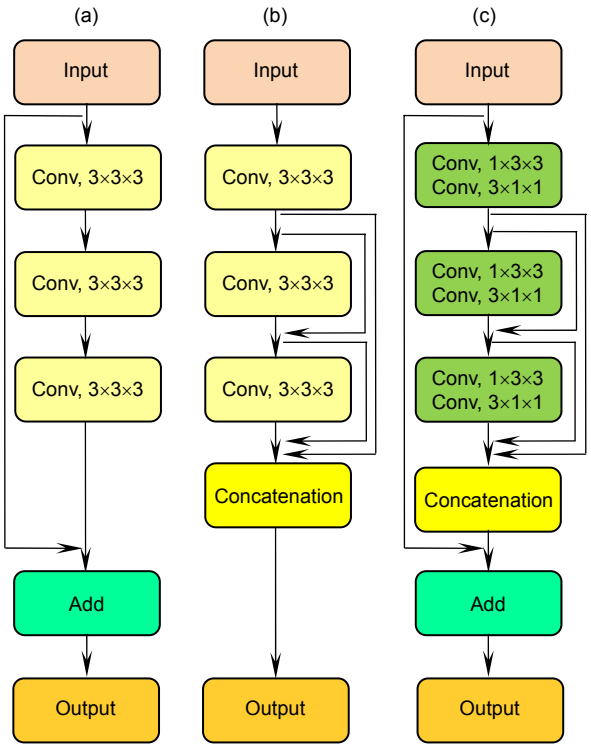


图 2 残差网络和密集连接。

(a) 残差块；(b) 密集块；(c) 密集连接残差块

Fig. 2 Residual network and dense connection. (a) Residual block; (b) Dense block; (c) Dense connection residual block

3 3D-EDRNs 网络及其在人体行为识别中的应用

现有 3D 卷积架构计算量大，对内存要求高，因此难以训练出较深的 3D 卷积架构。和计算需求很大的 3D 卷积相比，新提出的高效 3D 卷积块应用于经典 C3D 模型时，网络架构参数的数量大为降低，分类性能得到有效的提升。而且设计了一种融合改进的 3D 卷积块的高效密集残差网络，将 E3DB 放置在密集残差网络的合适位置，提高了网络的性能和效率。

3.1 提出的高效 3D 卷积块

当给定输入视频的尺寸为 $c\times l\times h\times w$，其中 c、l、h、w 分别为视频的通道数、视频长度、每帧图片的高度和宽度。3D 卷积可以像 2D 滤波器一样对空间信息进行建模，也可以构建每帧图片间的时序模型[18]。为简单起见，我们将三维卷积滤波器的大小表示为 $d\times k\times k$，d 是卷积核的时序深度，k 是卷积核的空间大小。根据双流网络两个分支同时获得时空特征的优势，将大小为 3×3×3 的 3D 卷积滤波器设计为 2D CNN 在空间域作用的 1×3×3 卷积滤波器和 1D CNN 在时间域作用的 3×1×1 卷积滤波器。卷积神经网络的空间复杂度决定了模型的参数数量，空间复杂度可以表示为

$$O\left(\sum_{l=1}^{D}K_l^2\times C_{l-1}\times C_l+\sum_{l=1}^{D}M_l^2\times C_l\right)\ , \tag{4}$$

其中 D 为 CNN 的卷积层数，l 为 CNN 第 l 个卷积层，M 为每个卷积核输出特征图的边长，K 为每个卷积核的边长，C_l 为第 l 个卷积层的输出通道数，$C_{l\text{-}1}$ 为第(l-1)个卷积层的输出通道数[19]。空间复杂度表示为总参数量和各层输出特征图的和，本文简化后忽略输出特征图对空间复杂度的影响。CNN 受到维度的限制，当模型的参数越多，训练模型所需的数据量就越大，而现实生活中的数据集不会很大，这样会导致模型训练时更容易过拟合。为简单起见，仅考虑卷积核的大小对空间复杂度的影响，E3DB 的空间复杂度是 3×3×3 卷积层的 $(K+1)/K^2$ (此处 K 取为 3)。

图 3(a)为标准的 3×3×3 卷积，E3DB 结构图如图 3(b)所示。这样的 3D 卷积是一种伪 3D 卷积块(E3DB)，E3DB 将参数量更少的时间域一维信息和空间域二维信息充分融合，空间维度卷积结果直接作为时间维度卷积的输入，有助于保留时空特征更加丰富的信息，从而可以减小模型的尺寸和提升网络的分类性能。

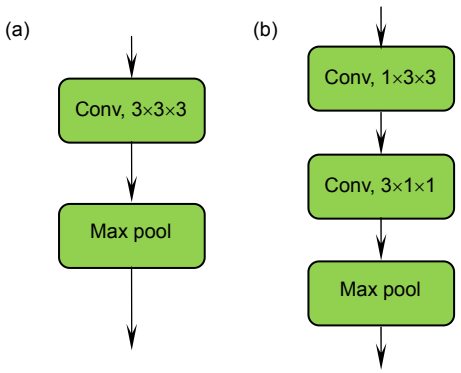


图 3 标准 3×3×3 卷积 (a)和 E3DB (b)

Fig. 3 Standard 3×3×3 convolution (a) and E3DB (b)

3.2 3D-EDRNs

C3D 网络在行为识别、场景识别、视频相似度分析等领域表现了较好的性能，该网络有 5 个 3×3×3 的卷积层。本文尝试将所提出的 E3DB 替换经典 C3D 网络中 3×3×3 卷积层，通过不同位置的 E3DB 组合测试

网络的综合性能。根据残差网络可以将信息流从浅层传输到较深层，密集网络可以将扩展的重复特性应用到整个网络，3D-EDRNs 设计为由一个小型的密集连接网络和一个残差结构组合而成，用于提取视频的时空域特征，网络的输入为一段连续视频。提出的 3D-EDRNs 结构如图 4 所示，其中 DRB 为密集连接残差块，结构如图 5 所示。为了确保每个 DRB 中 Add 层的输入都是未经过激活函数的特征图，将 DRB 内结构的连接顺序设置为融入 E3DB 的 3D 密集块、Add 层、3D 最大池化层、批量标准化(BN)、ReLU 激活函数。在 Concatenate 层之后，通过卷积层和池化层对输出的特征进行整合。

通过 DRB 的结构设计，3D-EDRNs 可以有效地获得网络卷积层间的信息流，有助于网络提取时空特征信息。Concatenate 层可以将网络所获取的浅层特征和高层特征进行充分融合。为了提升网络优化的速度和特征提取能力，3D-EDRNs 是一种残差网络和密集网络的高效融合形式。3D-EDRNs 提取视频中多变和复杂的时空特征，卷积层间的信息流也可以顺利传输到每一层，从而提高了网络参数的利用率，避免了普通深度神经网络参数膨胀的问题。

4 实验结果与分析

在本节中，为了评估提出的 E3DB 方法的有效性，在两个不同的视频分类网络中进行实验。首先在 4.1 节介绍了实验中所使用的 3 个人体行为数据库及数据预处理，4.2 节介绍了网络训练过程中参数的设置，4.3 节和 4.4 节分别叙述了 C3D 和 3D-EDRNs 两个不同视频分类网络的实验结果与分析。

4.1 数据集

UCF101 数据集就是其中的典型代表。UCF101 包含 13320 个视频(共 27 个小时)，每个视频中只包含一类人体行为，共有 101 个人体行为类别，例如运动、演奏、人与人互动和人物交互等，是目前行为类别数和样本数最多的数据库之一。国内外研究人员在 UCF101 数据集上进行了深入的研究，目前在此数据库上的准确率已达到 95%以上[20]。

首先，对训练的数据集进行预处理，将 UCF101 数据集中的所有视频保持结构不变，逐帧分解为图片保存到本地，然后制作图片的标签文档，跟 C3D 官方 Caffe 形式一致，模型输入的视频段长度为 16 帧，训练集与验证集的分割与 UCF101 官方保持一致，每个类别的前 7 个人为测试样本，后面 8~25 人为训练样本。

此外，本文还在 DMV action3D 数据库和 HMDB51[21]数据库上进行了实验。DMV action3D 数据库是多视角动态 3D 人体行为数据库，包括如鼓掌、自拍、喝水、读书和摔倒等 31 个不同的日常行为、交

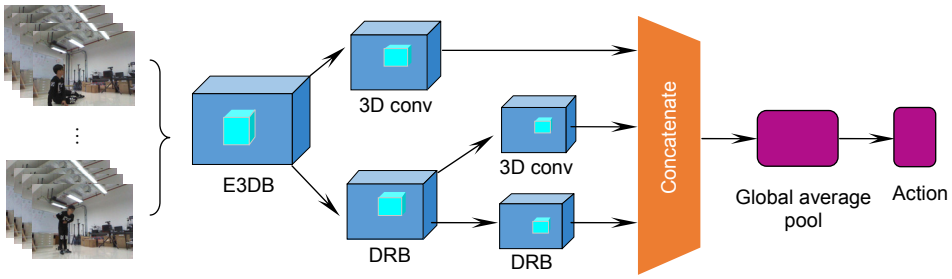


图 4 3D-EDRNs 结构图

Fig. 4 3D-EDRNs structure diagram

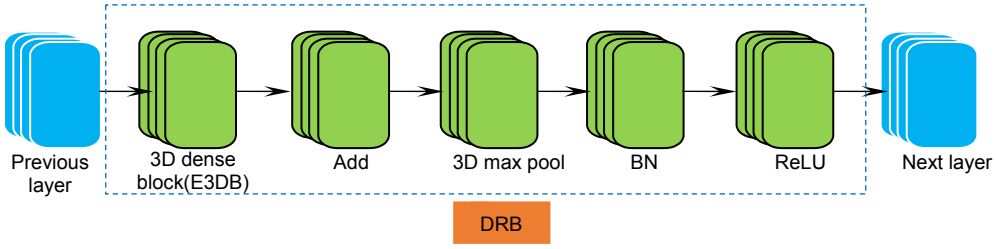


图 5 DRB 结构图

Fig. 5 DRB structure diagram

互行为和异常行为，为实验者分析视角和寻找最佳角度提供了可供验证的数据库。HMDB51 数据库从电影、YouTube 和谷歌视频中收集了 6849 个图像序列，分别代表 51 个行为类别，每个行为类至少包含 101 段样本，包括面部动作、与对象交互动作和身体动作等。

DMV action3D 数据库和 HMDB51 数据库训练前数据的预处理与 UCF101 数据集的处理方式一致。在训练集和测试集进行数据的预处理，目的是将每个样本置于更加规范的形式，以便减少模型需要考虑的变化量。减少数据中的变化量既能够提升模型的泛化能力，也能够减少拟合训练集所需模型的大小。

4.2 网络训练设置

在训练过程中首先将每一个输入的视频段(clip)中的每一帧的大小转换为 128×17，并在每个 clip 上裁剪出一个 112×112×16 大小的视频段作为模型输入(视频段长度为 16 帧)。在深度卷积神经网络的训练过程中通常会遇到过拟合的问题，即在训练集中损失函数值很小而在测试集中很大，数据增广技术是指对原图像进行各种变换，以增加样本的多样性，从而达到防止过拟合的目的，增强模型的鲁棒性[22]。本文对训练集中的每个 clip 都做一次翻转，作为简单的数据增广，可以减少模型的泛化误差，验证集只进行中心裁剪。为了评估网络架构的性能，在数据库上使用随机初始化的权值对网络模型进行从零开始的训练，网络训练权重衰减系数 weight_decay 设置为 0.005[23]。

4.3 C3D 网络融入 E3DB 的实验结果与分析

为了验证提出的高效 3D 卷积块的合理性和有效性，我们做了多个对比实验。经典 C3D 模型和融合 E3DB 的 C3D 模型在动态多视角复杂 3D 人体行为数据库进行的实验结果如表 1 所示。

从表 1 的实验结果可以看出：网络在 DMV action3D 数据库上提取人体行为时空特征时，不同视角的识别率有差异，说明了视角的不同对识别率有影响，其他视角出现部分肢体有遮挡、动作不完全等现象，导致动态视角的识别率最高。经典 C3D 模型融入 E3DB 后，每个视角的实验准确率都得到了提升(主视角提升 3.18%、侧视角提升 2.41%、动态视角提升 3.64%)。DMV action3D 数据库作为一个 3D 人体复杂行为数据库，可以提取的时空特征范围更广，C3D 模型融入 E3DB 可以更加有效地提取时空特征。经典 C3D 模型的参数量为 61.5 M，而 C3D 模型融入 E3DB 的参数量仅为 26.9 M，模型的参数量降低了一倍多，验证了 E3DB 可以大幅度降低网络的参数量，减少网络对计算和内存的需求(E3DB 的空间复杂度是 3×3×3 卷积层的$(K+1)/K^2$，此处 K 为 3)。E3DB 将参数量更少的时间域 1D CNN 信息和空间域 2D CNN 信息充分融合，空间维度卷积结果直接作为时间维度卷积的输入，有助于保留时空特征丰富的原始信息，从而获得了比参数量较大的 3D 卷积更好的效果。

4.4 3D-EDRNs 的实验结果与分析

我们将高效 3D 卷积块融合在密集残差网络的不同位置，得到了三种模型并进行对比实验，在 HMDB51 数据库上训练的准确率分别为 77.28%(底层特征融入 E3DB)、78.77%(底层特征和密集块融入 E3DB)、79.23%(高层特征、底层特征和密集块均融入 E3DB)。三种模型随迭代次数变化的准确率和损失值变化图分别如图 6~图 8 所示。3D-EDRNs 在 HMDB51 数据库实验的准确率如表 2 所示。

从图 6~图 8 的实验结果可以看出：3D-EDRNs 进行训练时，网络在迭代 15 次后开始收敛，三个不同位置组合高效 3D 卷积块的比较实验可以得到 E3DB 和密集块对提升网络性能都有显著的作用。当高层特征、底层特征和密集块均加入 E3DB 时，网络将所获取的浅层特征和高层特征进行充分融合，从而提升了网络优化的速度和特征提取能力，此时 3D-EDRNs 的分类性能最强，模型的准确率达到 79.23%。

从表 2 的实验结果可以看出，网络不同位置组合利用 E3DB 可以较好地提取出 HMDB51 数据库中的人体行为时空特征，有效地获得了卷积层间的信息流，

表 1 不同 C3D 模型基于 DMV action3D 数据库的实验结果

Table 1 Experimental results of different C3D models based on DMV action3D database

Method	Positive view recognition accuracy/%	Side view recognition accuracy/%	Dynamic view recognition accuracy/%	Model size/M
C3D	36.29	37.72	38.26	61.5
C3D+E3DB	39.47	40.13	41.9	26.9

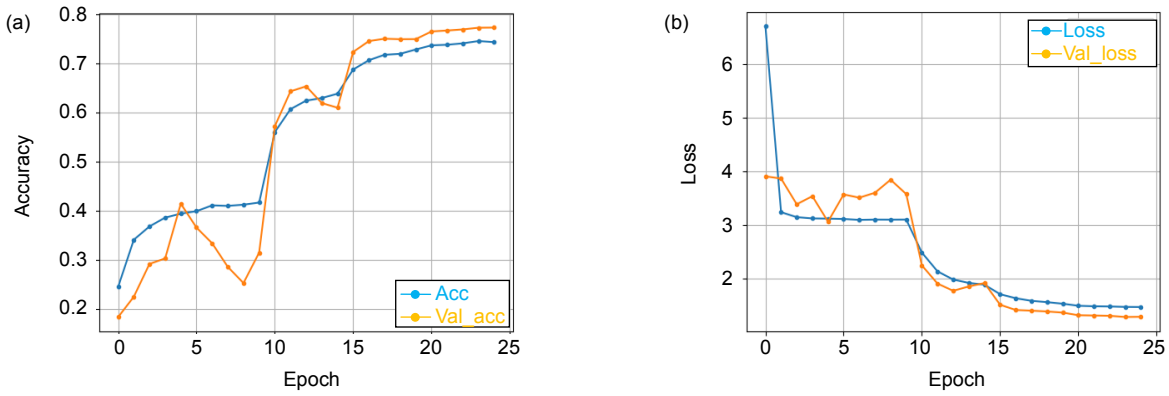


图 6 3D-EDRNs 在 HMDB51 的迭代准确率(a)和损失值(b)变化图(底层特征融入 E3DB)

Fig. 6 3D-EDRNs iteration accuracy (a) and loss value (b) variation diagram in HMDB51(lower features are integrated into E3DB)

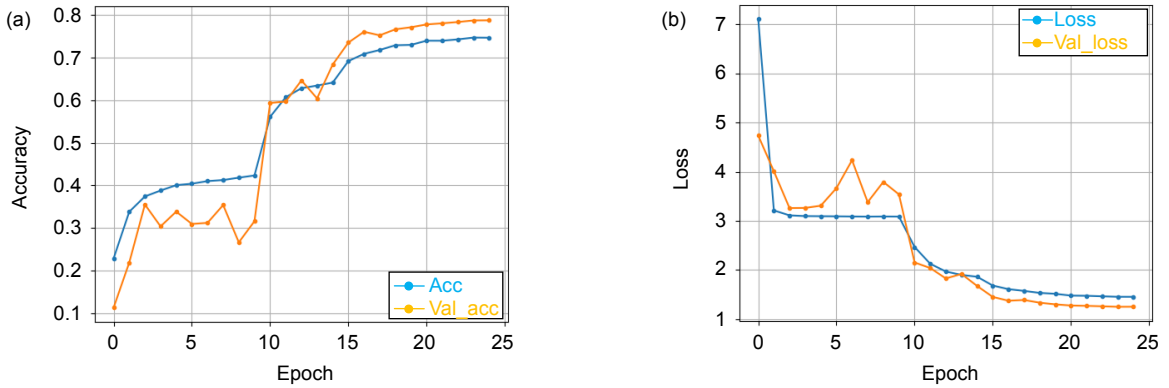


图 7 3D-EDRNs 在 HMDB51 的迭代准确率(a)和损失值(b)变化图(底层特征和密集块融入 E3DB)

Fig. 7 3D-EDRNs iteration accuracy (a) and loss value (b) variation diagram in HMDB51 (lower features and dense blocks are integrated into E3DB)

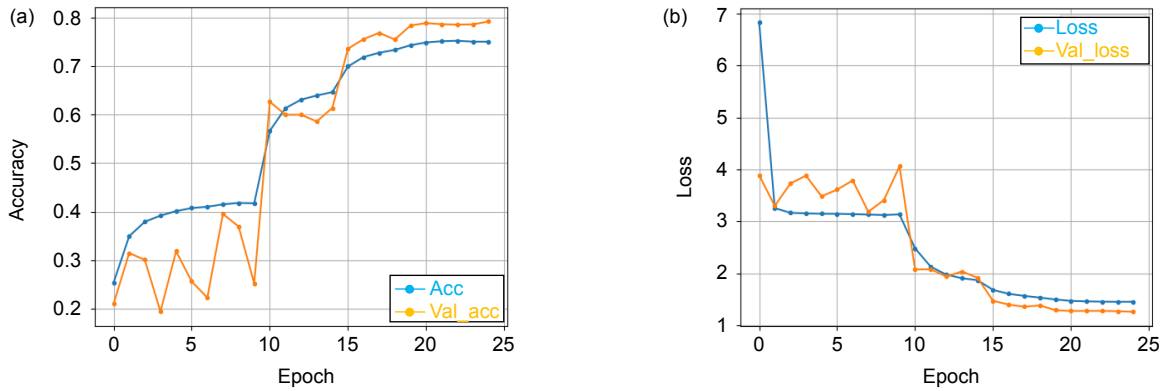


图 8 3D-EDRNs 在 HMDB51 的迭代准确率(a)和损失值(b)变化图(高层特征、底层特征和密集块均融入 E3DB)

Fig. 8 3D-EDRNs iteration accuracy (a) and loss value (b) variation diagram in HMDB51 (upper features, lower features and dense blocks are integrated into E3DB)

表 2 3D-EDRNs 基于 HMDB51 数据库的实验结果

Table 2 Experimental results of 3D-EDRNs based on HMDB51 database

Method	Accuracy/%	Model size/M
3D-EDRNs(before joining E3DB)+linear SVM	70.35	4.25
3D-EDRNs(lower features are integrated into E3DB)+linear SVM	73.26	4.6
3D-EDRNs(lower features and dense blocks are integrated into E3DB)+linear SVM	73.86	4.01
3D-EDRNs(upper features, lower features and dense blocks are integrated into E3DB)+linear SVM	76.29	3.97

很好地利用了 E3DB 能够降低网络参数量的特性，网络的参数量仅为 3.97 M，参数量反而比未加入 E3DB 的网络降低了 0.28 M，对计算机内存的使用率进一步减小。网络将参数量更少的时间域一维信息和空间域二维信息充分融合，空间维度卷积结果直接作为时间维度卷积的输入，保留了时空特征更加丰富的信息。随着 3D-EDRNs 加入 E3DB 和密集块，网络从包含丰富信息的视频中提取到多变和复杂的特征，分类性能逐步提高。3D-EDRNs 在高层特征、底层特征和密集块均加入 E3DB 时，网络优化的速度和特征提取能力达到最强，其准确率为 76.29%，比未加入 E3DB 的网络提升了 5.94%。3D-EDRNs 提取的特征输入线性支持向量机(linear SVM)进行实验时，准确率比在验证集上实验的结果略低，主要是由于 HMDB51 数据库人体行为类别较少，网络所提取的特征维度远远大于样本类别，此时网络分类加入 linear SVM 的效果会有所降低。

此外，3D-EDRNs 还在 UCF101 数据库上进行训练，得到三种位置的模型准确率分别为 58.18%(高层特征、底层特征和密集块均融入 E3DB)、55.95%(底层特征和密集块融入 E3DB)、55.07%(底层特征融入 E3DB)。三种模型随迭代次数变化的准确率变化图和损失值变化图分别如图 9~图 11 所示。3D-EDRNs 在 UCF101 数据库实验的准确率如表 3 所示。

从图 9~图 11 的实验结果可以看出：网络在迭代 15 次后均表现出较好的收敛性，3D-EDRNs 中所设计的 E3DB 和密集残差结构能够提升网络时空特征提取能力，网络的参数更易优化。3D-EDRNs 将从包含丰富信息的视频中获得的浅层特征和高层特征进行充分融合，提高了网络参数的利用率，其模型准确率达到 58.18%。

3D-EDRNs 提取的特征输入线性支持向量机(linear SVM)进行实验时，从表 3 的实验结果可以看出，随着网络加入 E3DB 和密集残差结构，网络从卷积层间获得了丰富的信息流，参数的利用率逐步提高。

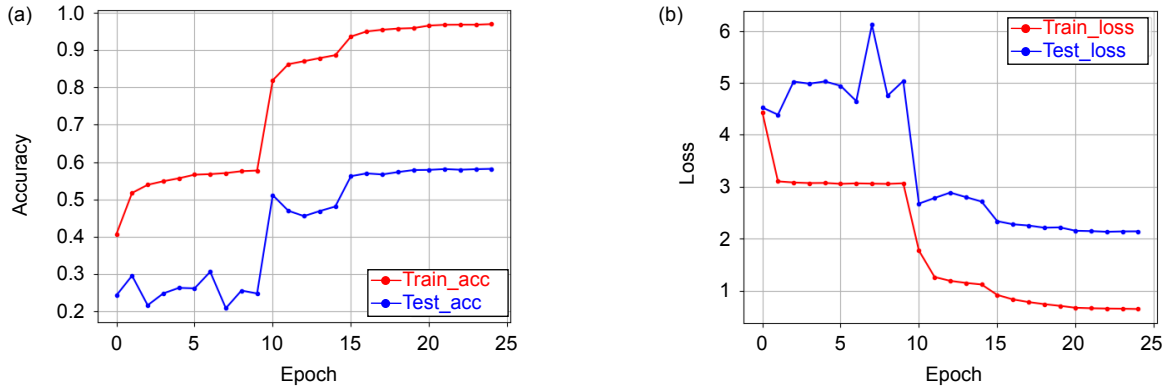


图 9 3D-EDRNs 迭代准确率(a)和损失值(b)变化图(高层特征、底层特征和密集块均融入 E3DB)

Fig. 9 Variation diagram of 3D-EDRNs iteration accuracy (a) and loss value (b) (upper features, lower features and dense blocks are integrated into E3DB)

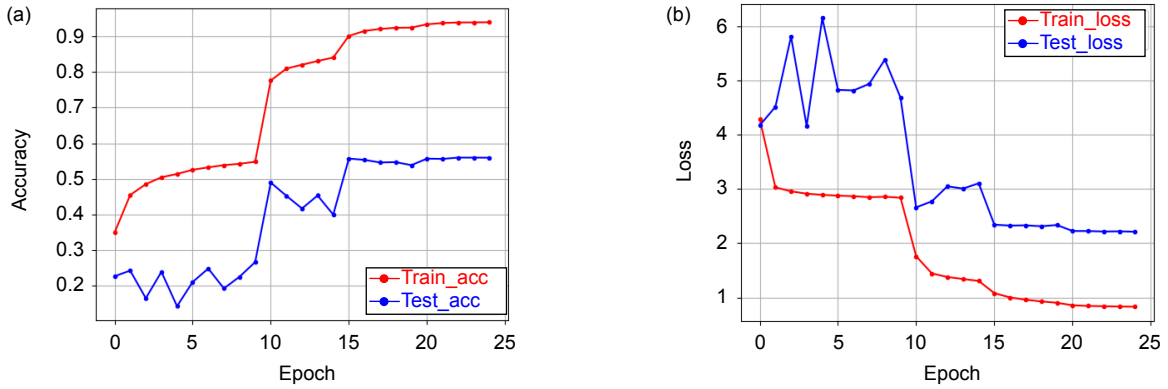


图 10 3D-EDRNs 迭代准确率(a)和损失值(b)变化图(底层特征和密集块融入 E3DB)

Fig. 10 Variation diagram of 3D-EDRNs iteration accuracy (a) and loss value (b) (lower features and dense blocks are integrated into E3DB)

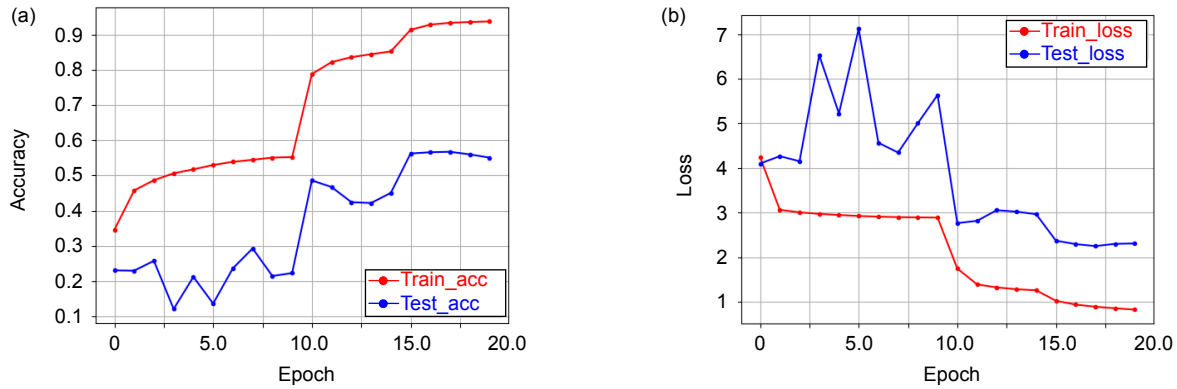


图 11 3D-EDRNs 迭代准确率(a)和损失值(b)变化图(底层特征融入 E3DB)

Fig. 11 Variation diagram of 3D-EDRNs iteration accuracy (a) and loss value (b) (lower features are integrated into E3DB)

表 3 3D-EDRNs 基于 UCF101 数据库的实验结果

Table 3 Experimental results of 3D-EDRNs based on UCF101 database

Method	Accuracy/%	Model size/M
3D-EDRNs(before joining E3DB)+linear SVM	88.3	3.84
3D-EDRNs(lower features are integrated into E3DB)+linear SVM	93.53	3.26
3D-EDRNs(lower features and dense blocks are integrated into E3DB)+linear SVM	94.06	2.66
3D-EDRNs (upper features, lower features and dense blocks are integrated into E3DB)+linear SVM	97.09	3.43

3D-EDRNs 在高层特征、底层特征和密集块均加入 E3DB 时，网络的参数得到高效利用，其准确率为 97.09%，比未加入 E3DB 的网络提升了 8.79%，网络的参数量仅为 3.43 M，反而比未加入 E3DB 的网络降低了 0.41 M。网络通过 E3DB 将参数量更少的时空信息充分融合，并保留了时空特征丰富的信息，3D-EDRNs 网络的参数量得到降低、分类性能得到提升，对计算资源的需求进一步减小。

为了验证所提出的 3D-EDRNs 具有较好的分类性能和对计算机内存需求小等优点，在 UCF101 数据库上，本文与其他视频特征提取方法做了对比实验，包括 C3D[16]、P3D[18]和 LTC[24]。如表 4 所示，可以看出 3D-EDRNs 在 UCF101 数据库提取了较好的视频特征，虽然 3D-EDRNs 网络结构因内部存在大量的跳跃连接而变得复杂，但准确率明显优于其他方法，3D-EDRNs 的准确率比经典 C3D 网络高 14.79%。当模型的参数量越小说明模型的计算量就越小，计算量越小则对内存的要求就越低，相应对计算资源的需求也就越少，3D-EDRNs 网络结合 E3DB，极大地降低了网络的参数量。3D-EDRNs 网络的参数量仅为 3.43 M，远远低于其他方法的参数量，只有 P3D 网络参数量的 1/28，极大地降低了对计算机内存的使用率。3D-EDRNs 网络预测一个视频段(16 帧)消耗时间为 11 ms，比 C3D 网络运行速度快一倍，在同等计算量的情况下，该网络可以显著降低对计算资源的需求，更快实现视频分类功能。3D-EDRNs 融合了密集网络和残差网络的特性，高效地利用了模型的浅层特征和高层特征，有助于提升网络提取时空特征的能力，从而进一步提高了网络的准确率。

3D-EDRNs 在 UCF101 和 HMDB51 两个数据库的

表 4 基于不同视频特征提取方法的实验结果(UCF101 数据库)

Table 4 Experimental results based on different video feature extraction methods (UCF101 database)

Method	Accuracy/%	Model size/M	Elapsed time/(ms/clip)
C3D+linear SVM	82.3	78.4	22.8
LTC+linear SVM	84.8	16.1	13.8
P3D ResNet+linear SVM	88.6	98	14.3
3D-EDRNs+linear SVM	97.09	3.43	11

实验结果均验证了网络具有较好的分类性能，并且可以降低网络对计算的需求，极大地减少了网络的参数量。在 HMDB51 数据库训练的实验效果优于 UCF101 数据库训练的实验效果，网络的收敛速度更快，其原因可能是 HMDB51 中样本的数据量更大，3D-EDRNs 网络可以提取到更加具有区分度的特征，从而提升了网络的分类性能。

5 结 语

本文提出的 3D-EDRNs 架构能够有效地学习视频中的时空特征，为基于视频的人体行为识别提供了一个有效的深度学习网络框架。特别地，提出的高效 3D 卷积块将参数量更少的时间域一维卷积信息和空间域二维卷积信息融合，替换原来计算需求大的 3×3×3 卷积层，具有显著降低模型大小的特性，网络的分类性能得到有效的提升。所提出的 3D-EDRNs 网络充分利用密集网络易于优化和残差网络特征复用的优点，提高了网络参数的利用率，避免了普通深度神经网络参数膨胀等问题，实验结果显示 3D-EDRNs 具备对计算资源需求少、计算速度快和网络参数量小等优点。

参考文献

[1] He K M, Zhang X Y, Ren S Q, *et al*. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//*2015 IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 2015: 1026–1034.

[2] Shojaeilangari S, Yau W Y, Li J, *et al*. Dynamic facial expression analysis based on extended spatio-temporal histogram of oriented gradients[J]. *International Journal of Biometrics*, 2014, **6**(1): 33–52.

[3] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition[C]//*Proceeding MM '07 Proceedings of the 15th ACM international conference on Multimedia*, New York, 2007: 357–360.

[4] Laptev I, Marszalek M, Schmid C, *et al*. Learning realistic human actions from movies[C]//*2008 IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 2008: 1–8.

[5] Willems G, Tuytelaars T, Van Gool L. An efficient dense and scale-invariant spatio-temporal interest point detector[C]//*European Conference on Computer Vision*, Berlin, 2008: 650–663.

[6] Wang H, Schmid C. Action recognition with improved trajectories[C]//*2013 IEEE International Conference on Computer Vision*, Sydney, 2014: 3551–3558.

[7] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//*Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, 2014: 568–576.

[8] Yao L, Torabi A, Cho K, *et al*. Describing videos by exploiting temporal structure[C]//*2015 IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 2015: 199–211.

[9] Shao L, Zhen X T, Tao D C, *et al*. Spatio-temporal laplacian pyramid coding for action recognition[J]. *IEEE Transactions on Cybernetics*, 2014, **44**(6): 817–827.

[10] Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition[C]//*2017 IEEE International Conference on Computer Vision Workshops* (*ICCVW*), Venice, 2017: 3154–3160.

[11] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]//*Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015: 448–456.

[12] Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 2017: 2261–2269.

[13] Song T Z, Song Y, Wang Y X, *et al*. Residual network with dense block[J]. *Journal of Electronic Imaging*, 2018, **27**(5): 053036.

[14] Wang Y X, Li X, Li L H. Dynamic and multi-view complicated 3D database of human activity and activity recognition[J]. *Journal of Data Acquisition & Processing*, 2019, **34**(1): 68–79.
王永雄, 李璇, 李梁华. 动态多视角复杂 3D 人体行为数据库及行为识别[J]. 数据采集与处理, 2019, **34**(1): 68–79.

[15] 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.

[16] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks[C]//*2015 IEEE International Conference on Computer Vision* (*ICCV*), Santiago, 2014: 4489–4497.

[17] Qiu Z F, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//*2017 IEEE International Conference on Computer Vision* (*ICCV*), Venice, 2017: 5534–5542.

[18] He K M, Sun J. Convolutional neural networks at constrained time cost[C]//*2015 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 2015: 5353–5360.

[19] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild[Z]. arXiv:1212.0402, 2012.

[20] Tran D, Torresani L. EXMOVES: mid-level features for efficient action recognition and video analysis[J]. *International Journal of Computer Vision*, 2016, **119**(3): 239–253.

[21] Wang Z L, Huang M, Zhu Q B, *et al*. The optical flow detection method of moving target using deep convolution neural network[J]. *Opto-Electronic Engineering*, 2018, **45**(8): 38–47.
王正来, 黄敏, 朱启兵, 等. 基于深度卷积神经网络的运动目标光流检测方法[J]. 光电工程, 2018, **45**(8): 38–47.

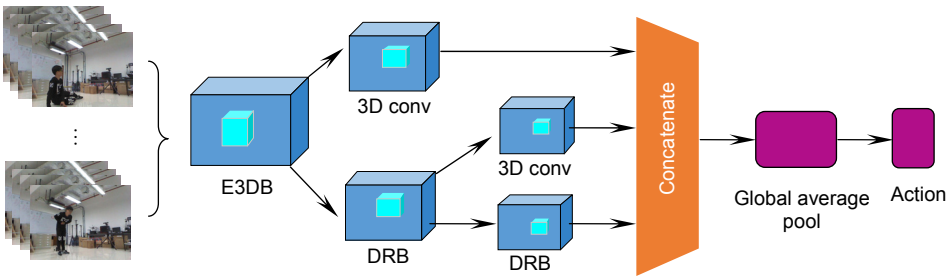
[22] Wang X H, Gao L L, Wang P, *et al*. Two-stream 3-D convNet fusion for action recognition in videos with arbitrary size and length[J]. *IEEE Transactions on Multimedia*, 2018, **20**(3): 634–644.

[23] 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.

Efficient 3D dense residual network and its application in human action recognition

Li Lianghua, Wang Yongxiong*

School of Optical-Electrical and Computer Engineering,
University of Shanghai for Science and Technology, Shanghai 200093, China



3D-EDRNs structure diagram

Overview: In view of the problem that 3D-CNN can better extract the spatio-temporal features in video, but it requires a high amount of computation and memory, this paper designs an efficient 3D convolutional block to replace the 3×3×3 convolutional layer with a high amount of computation, and then proposes a 3D-efficient dense residual networks (3D-EDRNs) integrating 3D convolutional blocks for human action recognition. The efficient 3D convolutional block is composed of 1×3×3 convolutional layers for obtaining spatial features of video and 3×1×1 convolutional layers for obtaining temporal features of video. The spatial dimension convolution results are directly used as the input of time dimension convolution, which is helpful to retain the original information with abundant spatio-temporal characteristics. According to the residual network, the information flow can be transmitted from the shallow layer to the deeper layer. The dense network can apply the extended repetition features to the entire network. 3D-EDRNs is designed as a combination of a small dense connection network and a residual structure, which is used to extract the spatial-temporal features of video. The new dense residual structure extends the original dense residual structure from 2D to 3D, and integrates E3DB, which can accelerate the network training and improve the performance of the residual network. Input of the add layer is processed through the structural design of the DRB, which are all feature graphs of inactivated functions, thus, 3D-EDRNs can effectively obtain the information flow between convolutional layers, which is helpful for the network to extract the spatial-temporal features. The concatenate layer can fully integrate the shallow and high level features obtained by the network. 3D-EDRNs extracts the variable and complex spatio-temporal features of video, and the information flow between convolutional layers can also be transmitted to each layer smoothly, thus improving the utilization rate of network parameters and avoiding the problem of parameter expansion of common neural networks. Efficient 3D convolutional blocks are combined in multiple locations of dense residual network, which not only takes advantage of easy optimization of residual blocks and feature reuse of dense connected network, but also can shorten the training time and improve the efficiency and performance of spatial-temporal feature extraction of the network. In the classical data set UCF101, HMDB51 and the dynamic multi-view complicated 3D database of human activity (DMV action3D), it is verified that the 3D-EDRNs combined with 3D convolutional block can significantly reduce the complexity of the model, effectively improve the classification performance of the network, and have the advantages of less computational resource demand, small number of parameters and short training time.

Citation: Li L H, Wang Y X. Efficient 3D dense residual network and its application in human action recognition[J]. *Opto-Electronic Engineering*, 2020, **47**(2): 190139

Supported by National Natural Science Foundation of China (61673276, 61603255, 61703277)
* E-mail: wyxiong@usst.edu.cn