智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种基于2D时空信息提取的行为识别算法

刘董经典，孟雪纯，张紫欣，杨旭，牛强

引用本文:
刘董经典, 孟雪纯, 张紫欣, 等. 一种基于2D时空信息提取的行为识别算法[J]. 智能系统学报, 2020, 15(5): 900–909.
LIU Dongjingdian, MENG Xuechun, ZHANG Zixin, et al. A behavioral recognition algorithm based on 2D spatiotemporal information extraction[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(5): 900–909.

在线阅读 View online: https://dx.doi.org/10.11992/tis.201906054

DOI: 10.11992/tis.201906054

一种基于 2D 时空信息提取的行为识别算法

刘董经典，孟雪纯，张紫欣，杨旭，牛强
（中国矿业大学 计算机科学与技术学院，江苏 徐州 221008）

摘　要：基于计算机视觉的人体行为识别技术是当前的研究热点，其在行为检测、视频监控等领域都有着广泛的应用价值。传统的行为识别方法，计算比较繁琐，时效性不高。深度学习的发展极大提高了行为识别算法准确性，但是此类方法和图像处理领域相比，效果上存在一定的差距。设计了一种基于 DenseNet 的新颖的行为识别算法，该算法以 DenseNet 做为网络的架构，通过 2D 卷积操作进行时空信息的学习，在视频中选取用于表征行为的帧，并将这些帧按时空次序组织到 RGB 空间上，传入网络中进行训练。在 UCF101 数据集上进行了大量实验，实验准确率可以达到 94.46%。

关键词：行为识别；视频分析；神经网络；深度学习；卷积神经网络；分类；时空特征提取；密集连接卷积网络
中图分类号：TP391.41　　**文献标志码：**A　　**文章编号：**1673−4785(2020)05−0900−10

中文引用格式：刘董经典，孟雪纯，张紫欣，等. 一种基于 2D 时空信息提取的行为识别算法 [J]. 智能系统学报，2020，15(5)：900–909.
英文引用格式：LIU Dongjingdian, MENG Xuechun, ZHANG Zixin, et al. A behavioral recognition algorithm based on 2D spatiotemporal information extraction[J]. CAAI transactions on intelligent systems, 2020, 15(5): 900–909.

A behavioral recognition algorithm based on 2D spatiotemporal information extraction

LIU Dongjingdian, MENG Xuechun, ZHANG Zixin, YANG Xu, NIU Qiang
(College of Computer Science & Technology, China University of Mining and Technology , Xuzhou 221008, China)

Abstract: Human behavior recognition technology based on computer vision is a research hotspot currently. It is widely applied in various fields of social life, such as behavioral detection, video surveillance, etc. Traditional behavior recognition methods are computationally cumbersome and time-sensitive. Therefore, the development of deep learning has greatly improved the accuracy of behavior recognition algorithms. However, compared with the field of image processing, there is a certain gap in the effect of such methods. We introduce a novel behavior recognition algorithm based on DenseNet, which uses DenseNet as the network architecture, learns spatio-temporal information through 2D convolution, selects frames for characterizing behavior in video, organizes these frames into RGB space in time-space order and inputs them into our network to train the network. We have carried out a large number experiments on the UCF101 dataset, and our method can reach an accuracy rate of 94.46%.

Keywords: behavior recognition; video analysis; neural networks; deep learning; convolutional neural networks; classification; spatiotemporal feature; densenet

近年来，人体行为识别技术随着深度学习的兴起，引起了广泛的关注。传统的行为识别方法，如 iDT[1]，计算繁琐，时效性不高。深度学习以及卷积神经网络的发展推动了行为识别技术的发展。主流深度学习网络模型，如 AlexNet[2]、VGG-Net[3]、GoogleLetNet[4]、ResNet[5] 和 DenseNet[6] 等，在 2D 图像数据处理方面取得了不错的效果。

基于深度学习的人体行为识别方法目前主要

收稿日期：2019−06−28.
基金项目：国家自然科学基金项目（51674255）.
通信作者：牛强. E-mail: .niuq@cumt.edu.cn.

包括两个流派：3D 时空卷积 (3D ConvNets) 和双流卷积网络 (Two-Stream)，主要基于的网络架构是 ResNet。

本文采用 DenseNet 做为网络的架构，通过 2D 卷积操作进行时空信息的学习，提出了一种新的基于视频的行为识别方法：2D 时空卷积密集连接神经网络 (2D spatiotemporal dense connected convolutional networks, 2DSDCN)。首先在视频中选取用于表征行为的帧，并将这些帧按时空次序组织成 BGR 格式数据，传入 2DSDCN 中进行识别。2DSDCN 模型在 DenseNet 的基础上添加了时空信息提取层，与单纯使用 DenseNet 相比，在 UCF101[7] 数据集上得到了 1% 的效果提升。目前，本文的方法在没有使用多流融合、iDT 信息融合等手段，在 UCF101 数据集上获得了最高 94.46% 的准确率。

本文提出了一种新的基于 2D 卷积的行为识别方法，使用 2D 卷积提取时空信息；引入了 DenseNet 作为行为识别的网络架构，分析其对时空信息提取的促进作用；提出了一种新的基于 BGR 图像的时空关系组织提取方法。

1　相关工作

1.1　卷积网络

卷积神经网络模型由交替堆叠的卷积层、池化层和全连接层构成。AlexNet、LeNet[8]、VGG-Net 在结构上并没有太大的改进，卷积层、池化层和全连接层进行合理的组织来加深模型的深度。GoogLeNet 引入了 Inception 结构来串联特征图，通过多分辨率来丰富提取到的特征。

ResNet 引入了残差块，即增加了把当前输出直接传输给后面层网络而绕过了非线性变换的直接连接，梯度可以直接流向前面层，有助于解决梯度消失和梯度爆炸问题。然而该网络的缺点是，前一层的输出与其卷积变换后的输出之间通过值相加操作结合在一起可能会阻碍网络中的信息流[5-6]。

DenseNet 在 ResNet 的基础上提出了一种不同的连接方式。它建立了一个密集块内前面层和后面所有层的密集连接，即每层的输入是其前面所有层的特征图，与 ResNet 在值上的累加不同，DenseNet 是维度上的累加，因此在信息流方面克服了 ResNet 的缺点，改进了信息流。DenseNet 的网络结构由密集块组成，其中，两个密集块之间有过渡层。密集块内的结构参照了 ResNet 的瓶颈结构 (Bottleneck)，而过渡层中包括了一个 1×1 的卷积层和一个 2×2 的平均池化层。DenseNet 减少了参数，使网络更窄，缓解了梯度消失问题，加强了特征的传播，鼓励特征重用[6]。

对于卷积网络而言，输入网络数据的宽度 (weight)、高度 (height)、通道数 (channels) 以及数据的分布对网络的实际表现有很大的影响。而这些卷积网络的源生输入数据均为 3 通道的 RGB 图像，数据未归一化前分布在 0~255。因此，为了充分发挥这些卷积网络的性能，本文决定将时空信息组织成 BGR 图像形式作为输入数据的组织形式。

1.2　行为识别算法

根据行为识别方法各自的特点，可大致分为基于特征工程的算法和基于深度学习的算法两大类。

基于特征工程的算法是传统的识别方法，其中最经典的是改进的密集轨迹算法[9-13](improved dense trajectories, iDT)。iDT 算法源于对 DT(dense trajectories) 算法的改进，主要思想是通过利用光流场来获得视频序列中的一些轨迹，再提取 HOF、HOG、MBH 等特征，用 BOF(bag of feature) 方法对提取到的特征进行编码，最后用 SVM 对编码的结果进行分类得到结果。iDT 在消除了相机运动带来的影响，优化了光流信息的同时，对提取的 HOF、HOG、MBH 等特征采用 L1 正则化后再对每个维度开方，并使用了费舍尔向量的编码方式对 DT 算法进行优化，在 UCF50 上的准确率从原本的 84.5% 提升到了 91.2%，在 HMDB51 上的准确率也从原本的 46.6% 提升到了 57.2%[1]。

基于深度学习的算法可分为基于卷积的行为识别算法[14]、基于 Two-Stream 架构的行为识别算法[15-18] 以及基于人体骨骼序列的行为识别算法[19-20]3 类。前两者对视频进行像素级别的识别，而后者则依赖于单帧关键点或骨架等信息进行时间上的识别。

基于卷积的行为识别算法，最经典的是 C3D。TRAN Du 等[21] 提出的 C3D (3D ConvNets) 的基本思想是将二维卷积拓展到三维空间，引入 3D 卷积提取时空特征。在 C3D 的启发下，一系列的 2D 卷积网络结构的 3D 卷积版本被用于行为识别，例如 3D ResNets[22]、P3D[23]、T3D[24] 等。为解

决 3D 卷积学习参数冗余导致学习困难，TRAN Du 与 WANG Heng 在 FSTCN (factorized spatio-temporal convolutional networks)[25] 的启发下提出了结合 2D 卷积和 3D 卷积的 $R(2+1)$D[26] 神经网络。(2+1)D 卷积核与 3D 卷积核对比如图 1 所示，$R(2+1)$D 神经网络将 3D 的时空卷积分解为了 2D 的空间卷积和 1D 的时间卷积，使得空间信息与时间信息分离开来，便于分别对时空信息进行优化。

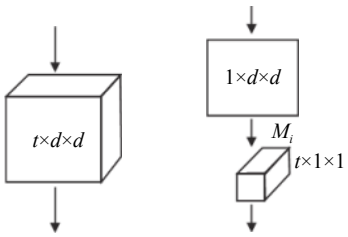


图 1　(2+1)D 卷积核与 3D 卷积核对比
Fig. 1　(2+1)D vs 3D convolution

基于 Two-Stream[27] 架构的行为识别算法通常对空间信息和时域信息进行分流学习然后将特征融合进行识别。比较经典的是 Simonyan 等[27] 提出的 Two-Stream Network。Two-Stream Network 训练了两个 CNN 学习，一个用于学习 2D 的 RGB 图，另一个用于学习光流信息，最后将两个分类器的结果融合起来。

基于人体骨骼序列的行为识别算法使用循环神经网络等方法，其通过时间序列上表征人体的关键点信息进行识别。现阶段主要利用的是骨架信息结合不同的循环神经网络进行研究。

现在主流的数据组织形式是 RGB 图像和光流图像。光流图像对运动的表征通常优于 RGB 图像。但是对光流的计算往往会带来时效上的损失，需要对新的数据组织形式进行探索。因此，本文尝试使用按照时间顺序组织的 RGB 平铺图像作为数据组织形式，通过 2D 卷积提取时空信息。

2　2D 时空卷积设计以及时空特征组织形式

本节对 2D 卷积用于时空特征提取的可能性进行分析，设计了适用于 2D 卷积的输入数据组织形式，分析了 DenseNet 在时空信息特征提取的促进作用，提出了最终的方案设计，如图 2 所示。

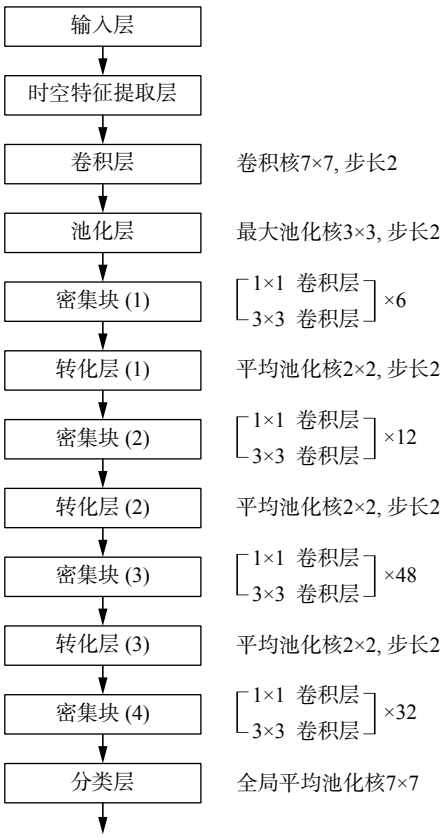


图 2　2DSDCN 网络架构
Fig. 2　Structure of 2DSDCN

2.1　2D 卷积理解与时空特征提取可行性分析

卷积神经网络 (CNN) 对信息特征的组织和提取主要依靠两种操作：卷积和池化操作。卷积依靠卷积集核将低层感受野中的相应信息组织到高层的对应像素点中。高层像素点 $A_{n,i,j}$ 以及对应的卷积核 C_n 中的信息表征了底层感受野中每个像素点 $A_{n-1,x,y}$ 和其他像素点之间的关系。随着卷积层数的加深，这种关系影响的范围根据卷积核的大小以线性级别扩大，最终在第 d 层的单个像素点中得到涵盖输入层每个像素点之间的关系，通过激活函数 φ 以及反向传播的方式可以拟合出输入层每个像素点 $A_{0,i,j}$ 和其他像素点之间的一个非线性的函数关系 R。对于一个输入、输出矩阵大小均为 w、卷积核大小为 k 的连续卷积操作来说，d 满足式 (1)：

$$d=\frac{2\times w}{k} \tag{1}$$

若在卷积过程中，在合适位置使用 n 次过滤窗口大小和移动步长均为 f 的池化操作，d 满足式 (2)：

$$d=\frac{2\times w}{k\times f^n} \tag{2}$$

可以看出池化操作对底层像素点之间关系的建立起到了不错的加速效果，使模型可以在尽量少的层次中获取对输入图像的表征。

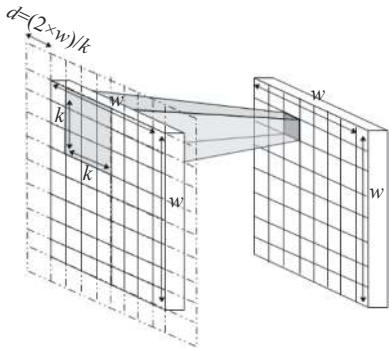
对一个卷积核尺寸为 k 的第 n 次卷积的输出层来说，第 i 行第 j 列的像素点 $A_{n,i,j}$ 代表了第 $n-1$ 次卷积的输出层 A_{n-1} 中，部分像素点之间的一个非线性关系 R：

$$A_{n,i,j}=R\begin{bmatrix} A_{n-1,i-\frac{k-1}{2},j-\frac{k-1}{2}} & \cdots & A_{n-1,i-\frac{k-1}{2},j+\frac{k-1}{2}} \\ \vdots & & \vdots \\ A_{n-1,i+\frac{k-1}{2},j-\frac{k-1}{2}} & \cdots & A_{n-1,i+\frac{k-1}{2},j+\frac{k-1}{2}} \end{bmatrix} \tag{3}$$

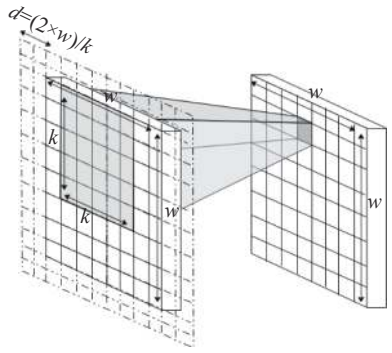
随着进一步卷积，$A_{n,i,j}$ 在 A_0 层所能表征的范围大小 r_n 会以 $A_{0,i,j}$ 为中心按图 3 所示方式扩大。其中，

$$r_n=(r_{n-1}+k-1)\times f_{n-1} \tag{4}$$

式中：f_{n-1} 表示在第 $n-1$ 次卷积后进行池化的步长，无池化操作时 $f_{n-1}=1$。



(a) k=3



(b) k=5

图 3　不同卷积核的卷积对比

Fig. 3　Comparison of different convolution kernels

对于单帧图像而言，2D 维度上的卷积可以提取到丰富的空间特征，这种特征是由单帧图像每个像素点与其他像素点之间的关系来进行表征。本文将多帧在时间上有相互关系的图像组织到空间维度上，通过 2D 卷积建立起同帧像素点间以及跨帧像素点间的关系，就可以提取到空间 (同帧内部) 和时间 (多帧之间) 上的关系。据此，本文提出了基于 RGB 图像和 2D 卷积对视频时空信息提取的方法。

2.2　选取和拼接的组织

本文从一个视频片段提取出 16 帧的 64×64×3 的有时序顺序的图像，组织成 4×4 的图像，组织的顺序如图 4 所示，最终的维度为 256×256×3。该图像的组织方法基于实际运算中卷积核的滑动的顺序，如图 5 所示，在卷积运算时，卷积核在相邻像素点之间建立关系，这可以保证每一帧都是先和相邻时域之间建立联系。卷积核横向移动时可以学习到大粒度的动作特征，纵向移动时则可以学习连续帧之间动作特征，丰富了时间特征的维度。随着卷积层数的增加，最终通过网络拟合出整个时空域之间的关系。

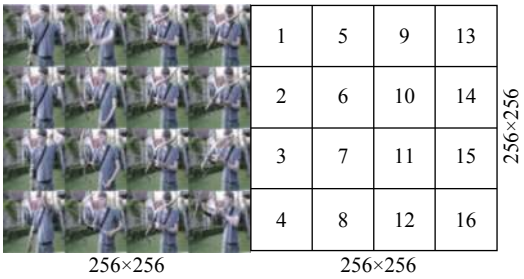


图 4　图像拼接

Fig. 4　Image mosaicking

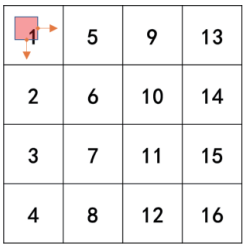


图 5　卷积示意图

Fig. 5　Convolution diagram

2.3　翻转操作及原因

单纯的拼接虽然可以快速提取相邻帧之间的关系，但是在建立不同帧中相邻空间像素点之间关系时，2D 卷积相比 3D 卷积有一定的差距：如图 6 所示，只有当 r_n 的大小超过 64 时，相邻空间像素点之间关系才会开始被建立起来。

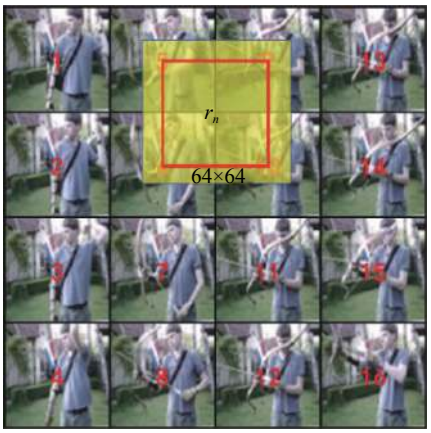
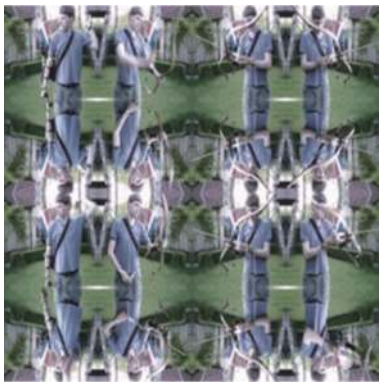


图 6　单纯拼接的缺点

Fig. 6　Disadvantages of simple mosaicking

为加快不同帧中相邻空间像素点之间关系的建立，本文对 16 帧图像进行翻转的操作，如图 7 所示，其中 H 代表水平翻转，V 代表垂直翻转。



256×256
(a) 翻转拼接图像

1	5	9	13
2	6	10	14
3	7	11	15
4	8	12	16

256×256
(b) 视频帧顺序

0	H	0	H
V	HV	V	HV
0	H	0	H
V	HV	V	HV

256×256
(c) 翻转操作示意

256×256

图 7　图像翻转设计

Fig. 7　Image reversal design

在卷积核位于多帧交界处时，能够在首次卷积中就建立范围 r_0 间像素点之间的关系。随着卷积核的移动，该范围不会只影响单帧图像的边缘，随着深度 d 的加深，当 r 的大小变为 64 时，覆盖了所有单帧空间上的像素点，过程如图 8 所示。

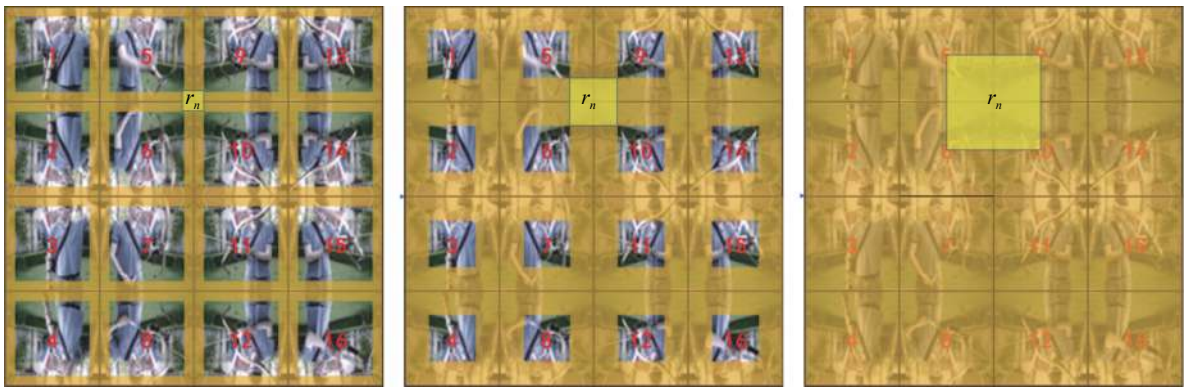


图 8　像素点覆盖

Fig. 8　Pixel point coverage

通过此操作可以加快部分相邻帧间对应像素点之间联系的提取，使相邻帧图像帧之间的时空联系在更低的层次上建立起来。对比无翻转的组织形式，能够在相同深度下更好地提取时空信息。

2.4　DenseNet 的选择

DenseNet 是 CVPR2017 的最佳论文，不同于之前的神经网络在宽度 (inception 结构) 和深度 (resblock 结构) 上的改进，在模型的特征维度进行了改进，将不同卷积阶段所提取的特征进行维度上的密集连接，可以保留更丰富的信息。DenseNet 建立了一个 denseblock 内前面层和后面所有层的密集连接，即每层的输入是其前面所有层的特征图，第 l 层的输出 x_l 可以表示为如下恒等函数：

$$x_l = H_l(\{x_0, x_1, \cdots, x_{l-1}\}) \tag{5}$$

式 (5) 中：$\{x_0, x_1, \cdots, x_{l-1}\}$ 表示 $0 \sim l-1$ 层的输出特征图的集合，在运算时按照通道的维度拼接在一起，作为第 l 层的输入。

对本文提出的方法来说，不同卷积阶段所提取的特征 x_l 表征了不同时空维度的信息，经过 DenseNet 架构训练，对 x_l 进行卷积运算意味着对所有已得的特征 x 都进行了进一步的特征提取，可以很好地保留不同维度的时空信息，更有利于时空特征的提取。

2.5　引入时空卷积层提取时空信息

结合 2.1 节提出加入 4 个大小为 33×33 的大

卷积核作为时空卷积层预提取层，每个卷积层后加入 BN 层，仿照 resnet 的 resblock 以及 DenseNet 的 densblock 思想，本文设计了两种结构，如图 9 所示，由式 (4) 可知，此时 $r_4 = 128$。在时空卷积层之后使用 DenseNet 201 作为模型的 baseline 进行分类的训练。本文采用了与 keras 内置版本不同，在每次 BN 操作之后加入 Scale 层操作。连续的 4 次 33×33 的卷积操作，能够快速建立输入数据的时空关系，使得 baseline 网络能够对更高层次的时空关系进行学习。最终结构如图 2 所示，结合图 9 可得 2 个不同的网络：2DSDCN_R 和 2DSDCN_D。

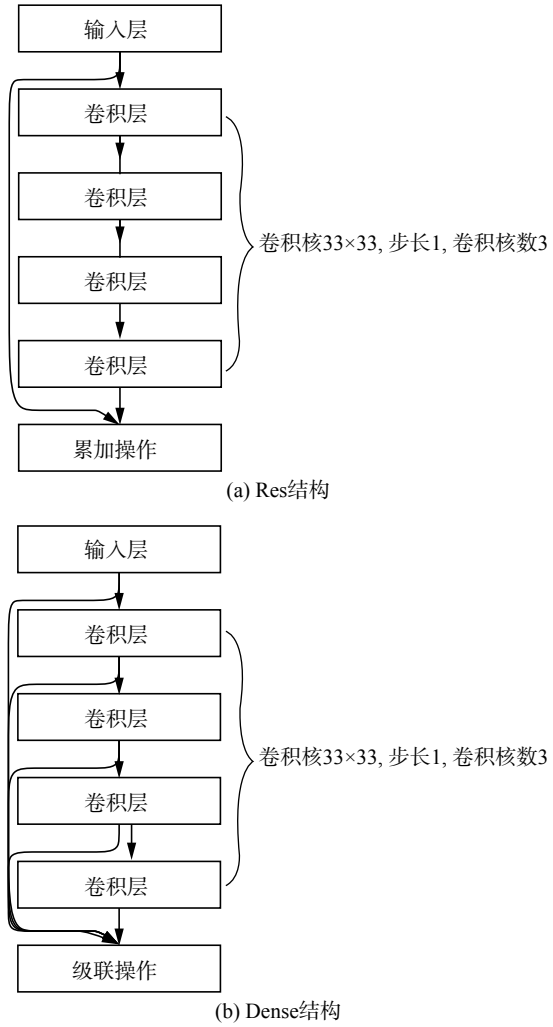


图 9　时空卷积层结构

Fig. 9　Structure of spatiotemporal convolutional layer

3　实验

本文首先通过对比实验来验证 2DSDCN 各个部分的设计，然后对输入视频采用不同的帧选取方式，来验证模型的鲁棒性。实验所采用的数据集为 UCF101 数据集，选用以 tensorflow 为后端的 Keras 框架，在训练过程中使用 2 个单精度 GPU 进行加速，型号为 Pascal 架构下的 GTX 1080Ti。

3.1　翻转操作的验证

本文对每一个视频等分采样 16 帧，设单个视频的总帧数为 Fl，对应的采样帧间隔为 Fl/16。将得到的 16 帧图像进行翻转或拼接操作，分别得到直接拼接的 BGR 图像和使用翻转拼接的 BGR 图像。视频数据集按照相同的随机系数进行打乱，训练集和验证集按照 8:2 的比例划分。本文将获得的数据集直接送入没有进行预训练的 DenseNet-201 网络进行训练，训练轮次为 100 轮。训练的结果如图 10 所示。

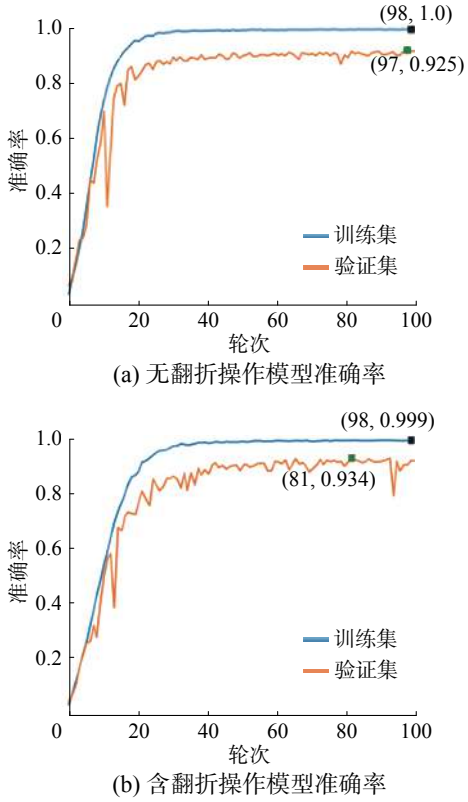


图 10　无翻转操作与带翻转操作准确率对比

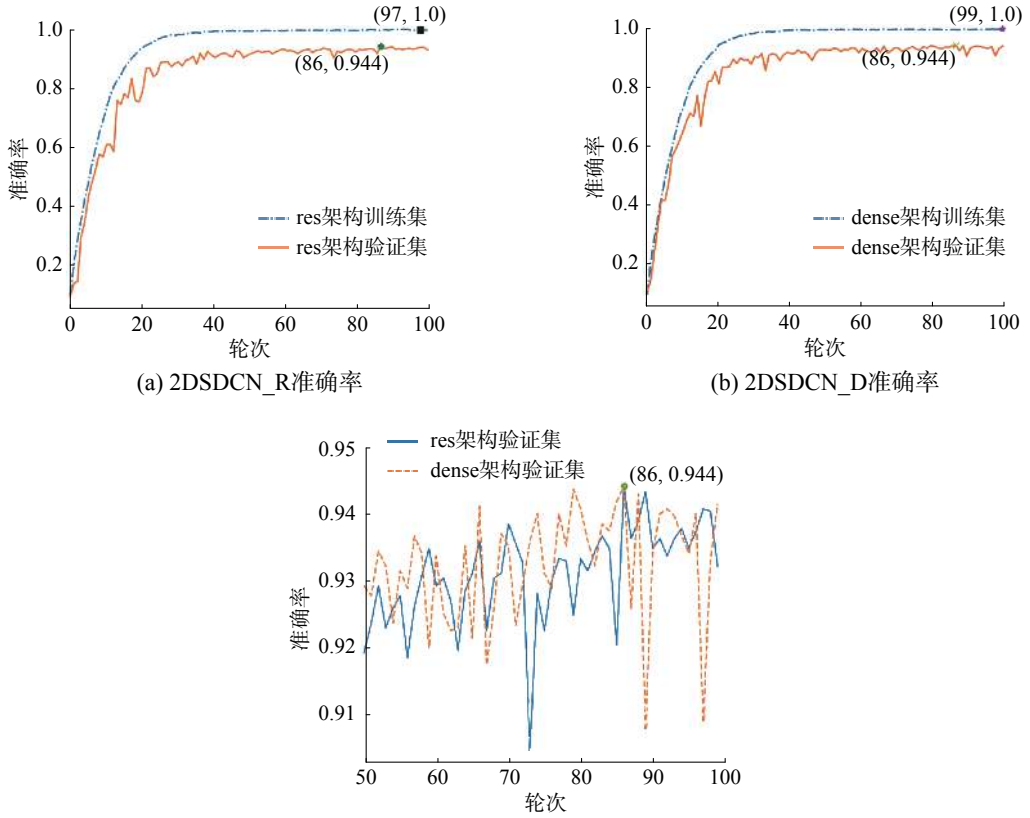
Fig. 10　Accuracy comparison between no flipping operation and flipping operation

没有进行翻转操作的模型训练后的准确率为 92.5%，而带翻转操作的模型训练后的准确率则为 93.4%，有 1% 的效果提升，说明加快相邻帧间对应像素点之间关系的建立对模型学习时空信息起到了一定的促进作用。

3.2　时空卷积层效果提升与特征可视化

本文基于 resblock 和 denseblock 设计实现了 2 种不同的网络作为时空卷积层，并将翻转拼接图像数据集的 BGR 形式分别送入两个网络进行

训练，训练轮次为 100 轮，训练结果如图 11 所示。可以看出，二者在大卷积核操作的促进下，效果差距不大，均能达到 94.4% 的准确率，并且相比原始的 DenseNet-201 网络，有 1% 的效果提升。为了直观展现时空卷积层的效果，本文进行了特征的可视化，对每一层的输出，如图 12 所示。

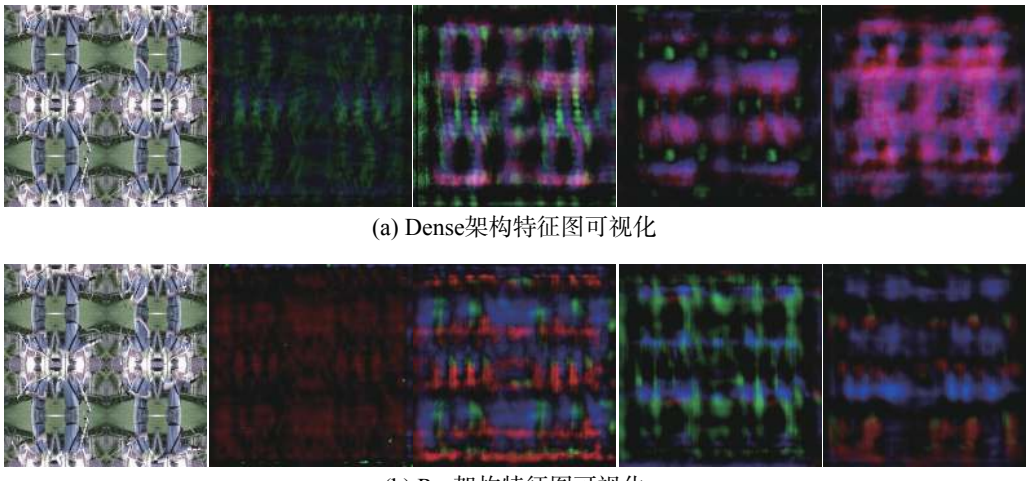


(a) 2DSDCN_R准确率　(b) 2DSDCN_D准确率

(c) 2DSDCN_R与2DSDCN_D对比

图 11　2DSDCN_R 和 2DSDCN_D 的准确率对比

Fig. 11　Accuracy comparison between 2DSDCN_R and 2DSDCN_D



(a) Dense架构特征图可视化

(b) Rse架构特征图可视化

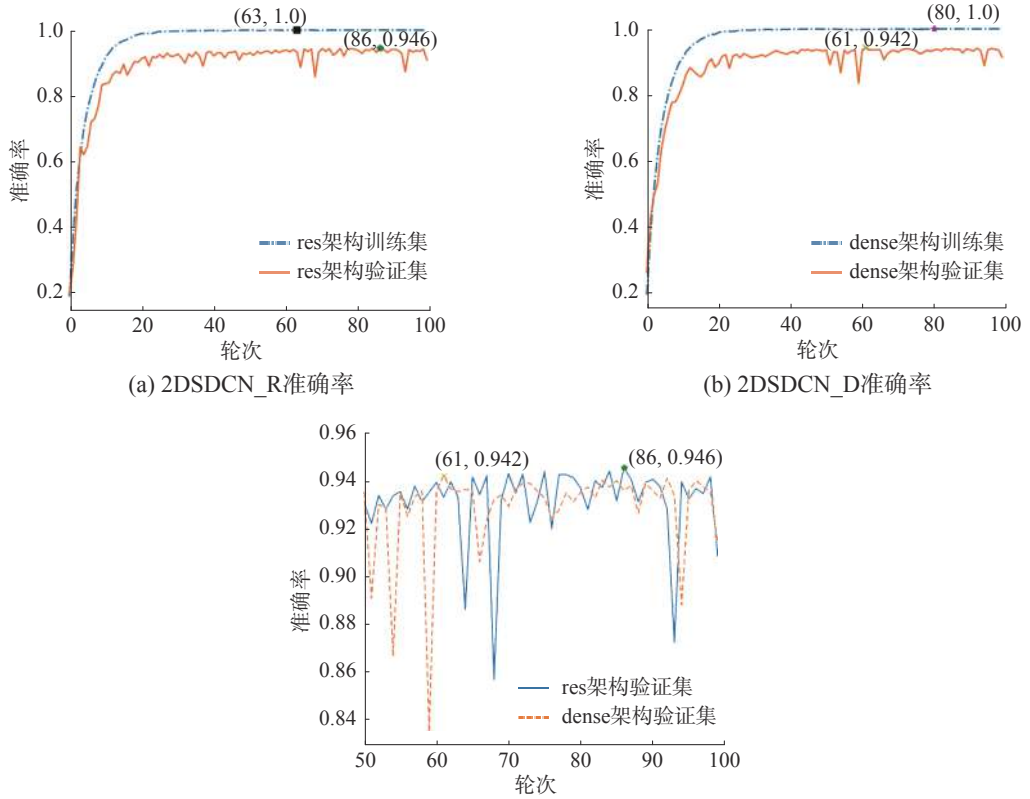
图 12　denseblock 和 resblock 设计的特征可视化

Fig. 12　Feature of visualization denseblock and resblock

3.3　不同的帧选取方式下模型鲁棒性的验证

在之前的帧选取方法的基础上，本文又对同一视频片段每隔 5 帧进行 16 帧的选取，对每一个视频进行随机划分，送入 3.2 节设计的两个网络中进行训练。之所以这样选取是因为在实际使用中，视频的获取方式往往是连续的。采用这种方

式获取的动作信息在时序上是等时分布的，符合实时识别的数据采样形式。训练结果如图 13 所示，可以发现模型准确率在 denseblock 设计上达到了 94.2%，在 resblock 设计上达到了 94.6%。



(a) 2DSDCN_R准确率　　(b) 2DSDCN_D准确率

(c) 2DSDCN_R与2DSDCN_D对比

图 13　每 5 帧采样下 2DSDCN_R 和 2DSDCN_D 的准确率对比

Fig. 13　Accuracy comparison between 2DSDCN_R and 2DSDCN_D with sampling every 5 frames

3.4　实验分析

本文对之前的实验进行汇总，详细内容参见表 1。2DSDCN 结合翻转平铺的数据组织形式对比无附加操作时，准确率更高，收敛速度更快。本文所设计的两种时空特征层提取结构在实验中均能达到相同的水准，并且在不同的视频数据采样形式下保持稳定的准确率，表示基于 2D 时空平展图的大卷积核时空特征提取能够有效加速时空关系的建立。

表 1　实验结果

Table 1　Experiment result

模型	拼接方式	时空卷积层	帧选取间隔	最佳轮次	准确率/%
DenseNet-201	无翻转	无	Fl/16	97	92.5
DenseNet-201	含翻转	无	Fl/16	81	93.4
2DSDCN_R	含翻转	Resblock	Fl/16	86	94.4
2DSDCN_D	含翻转	Denseblock	Fl/16	86	94.4
2DSDCN_R	含翻转	Resblock	5	86	94.6
2DSDCN_D	含翻转	Denseblock	5	61	94.2

4　结束语

本文在对过往行为识别特别是基于卷积的行为识别算法总结分析的基础上，提出了一种新的基于视频的 2D 卷积行为识别算法 2DSDCN 以及一种新的时空信息数据的组织形式，并对 2D 卷积在像素级别上进行分析，引入 dense 结构来进行时空信息的提取。

本文以基础 CNN 分类网络对 BGR 图像组织形式的良好支持为源起，通过合理的组织输入数据，结合 CNN 网络的特性，最终设计出两种性能稳定的网络 2DSDCN_R 和 2DSDCN_D，并在 UCF101 数据集上取到了最高 94.6% 的准确率。

图像识别和视频行为识别领域存在着一定的技术隔阂，本文尝试通过合理的组织数据将图像识别和视频识别有机统一，一定程度上促进了图像识别与视频识别的同步发展，但还需要进一步的研究，以支持视频分割以及视频语义理解等研究方向。

参考文献：

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

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097−1105.

[3] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International journal of computer vision, 2014, 115(3): 211–252.

[4] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1−9.

[5] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770−778.

[6] HUANG G, LIN Z, LAURENS V D M, et al. Densely connected convolutional networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 2261−2269.

[7] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv: 1212.0402, 2012.

[8] 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.

[9] CHEN P H, LIN C J, Schölkopf B. A tutorial on v-support vector machines[J]. Applied stochastic models in business and industry, 2005, 21(2): 111–136.

[10] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005, 1: 886−893.

[11] CHAUDHRY R, RAVICHANDRAN A, HAGER G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA, 2009: 1932−1939.

[12] DALAL N, TRIGGS B, SCHMID C. Human detection using oriented histograms of flow and appearance[C]//European Conference on Computer Vision. Graz, Austria, 2006: 428−441.

[13] WANG H, Kläser A, SCHMID C, et al. Action recognition by dense trajectories[C]//Proceedings of the IEEE International Conference on Computer Vision. Colorado Springs, USA, 2011: 3169−3176.

[14] 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. Venice, Italy, 2017: 6299−6308.

[15] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1933−1941.

[16] NG Y H, HAUSKENCHT 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.

[17] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 20−36.

[18] LAN Z, ZHU Y, HAUPTMANN A G, et al. Deep local video feature for action recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Venice, Italy, 2017: 1−7.

[19] 张培浩. 基于姿态估计的行为识别方法研究 [D]. 南京: 南京航空航天大学, 2015.
ZHANG Peihao. Research on action recognition based on pose estimation[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2015.

[20] 马淼. 视频中人体姿态估计、跟踪与行为识别研究 [D].

山东: 山东大学, 2017.
MA Miao. Study on human pose estimation, tracking and human action recognition in videos[D]. Shandong: Shandong University, 2017.

[21] TRAN D, BOURDEY 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.

[22] HARA K, KATAOKA H, SATOH Y. Learning spatio-temporal features with 3D residual networks for action recognition[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, Italy, 2017: 3154−3160.

[23] 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.

[24] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3d convnets: new architecture and transfer learning for video classification[J]. arXiv preprint arXiv: 1711.08200, 2017.

[25] SUN L, JIA K, YEUNG D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4597−4605.

[26] 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, USA, 2018: 6450−6459.

[27] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 568−576.

作者简介：

刘董经典，博士研究生，主要研究方向为行为识别、计算机视觉。

张紫欣，硕士研究生，主要研究方向为行为识别、推荐系统、智慧医疗。

牛强，教授，主要研究方向为人工智能、数据挖掘和无线传感器网络。发表学术论文 40 余篇。

[责任编辑：李雪莲]