·开发研究与工程应用· 文章编号：1000-3428(2021)11-0313-08 文献标志码：A 中图分类号：TP391

基于注意力机制的3D DenseNet人体动作识别方法

张聪聪[1]，何 宁[2]，孙琪翔[1]，尹晓杰[2]
（1.北京联合大学 北京市信息服务工程重点实验室，北京 100101；2.北京联合大学 智慧城市学院，北京 100101）

摘 要：传统人体动作识别算法无法充分利用视频中人体动作的时空信息，且识别准确率较低。提出一种新的三维密集卷积网络人体动作识别方法。将双流网络作为基本框架，在空间网络中运用添加注意力机制的三维密集网络提取视频中动作的表观信息特征，结合时间网络对连续视频序列运动光流的运动信息进行特征提取，经过时空特征和分类层的融合后得到最终的动作识别结果。同时为更准确地提取特征并对时空网络之间的相互作用进行建模，在双流网络之间加入跨流连接对时空网络进行卷积层的特征融合。在UCF101和HMDB51数据集上的实验结果表明，该模型识别准确率分别为94.52%和69.64%，能够充分利用视频中的时空信息，并提取运动的关键信息。

关键词：动作识别；注意力机制；三维DenseNet；双流网络；特征融合

开放科学（资源服务）标志码（OSID）：

中文引用格式：张聪聪，何宁，孙琪翔，等.基于注意力机制的3D DenseNet人体动作识别方法[J].计算机工程，2021，47(11)：313-320.

英文引用格式：ZHANG C C，HE N，SUN Q X，et al.Human motion recognition method based on attention mechanism of 3D DenseNet[J].Computer Engineering，2021，47(11)：313-320.

Human Motion Recognition Method Based on Attention Mechanism of 3D DenseNet

ZHANG Congcong[1]，HE Ning[2]，SUN Qixiang[1]，YIN Xiaojie[2]
（1.Beijing Key Laboratory of Information Service Engineering，Beijing Union University，Beijing 100101，China；
2.Smart City College，Beijing Union University，Beijing 100101，China）

【Abstract】Traditional human motion recognition algorithms cannot fully utilize the spatial and temporal information of human motions in videos，and are limited in the recognition accuracy.To address the problem，a three-dimensional dense convolutional network is proposed for recognizing human motions in videos.The model takes the two-stream network as the basic framework，using the 3D dense network with an attention mechanism for the spatial network to extract the apparent information features of human motions in the videos.On this basis，the temporal network is also used to extract the motion information features of the optical flows in the continuous video sequence.Then the spatio-temporal features and the classification layer are fused to obtain the final motion recognition accuracy.To extract features more accurately and model the interactions between the spatio-temporal networks，cross-stream connections are added between the two-stream networks to fuse the features at the convolutional layer of spatio-temporal networks.The experimental results show that the proposed model exhibits a recognition accuracy of 94.52% on the UCF101 dataset and 69.64% on the HMDB51 dataset.The model can make full use of the spatio-temporal information in the video to extract the key information of motions.

【Key words】motion recognition；attention mechanism；3D DenseNet；two-stream network；feature fusion

DOI：10.19678/j.issn.1000-3428.0059640

基金项目：国家自然科学基金（61872042，61572077）；北京市教委科技重点项目（KZ201911417048）；北京联合大学人才强校优选计划（BPHR2020AZ01，BPHR2020EZ01）；北京联合大学研究生科研创新项目（YZ2020K001）；“十三五”时期北京市属高校高水平教师队伍建设支持计划（CIT&TCD 201704069）。

作者简介：张聪聪（1995—），男，硕士研究生，主研方向为数字深度学习、图像处理；何 宁（通信作者），教授、博士；孙琪翔、尹晓杰，硕士研究生。

收稿日期：2020-10-24 **修回日期：**2020-11-24 **E-mail：**zhangcongc20@163.com

计算机工程 www.ecice06.com

0 概述

人体动作识别已成为计算机视觉领域的研究热点之一，在视频监控[1]、人机交互、医疗保健[2]、智能人机界面[3]等领域具有广泛的应用[4]。不同于图像处理，视频中动作识别仍然是一个具有挑战性的问题。视频序列包括空间特征和时序特征，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类特征。有效地从人体动作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，视频由大量的连续帧序列组成，具有极大的变化性和复杂性，例如遮挡、视点变化、背景差异、混乱、照明变化等情况[5]，这对视频中人体动作的识别提出了更高的要求。

随着计算机视觉的发展，卷积神经网络（CNN）已经从图像分类问题扩展到视频中动作识别问题[6]。在针对视频中动作的识别问题中，多数研究都集中在提取可以表达视频动作的有效特征上，与图像空间中的特征表示不同，视频中人体动作的特征表示不仅要描述图像空间中人的表观特征，而且还需提取动作外观和姿势的变化。因此，特征表示的问题从二维空间扩展到了三维时空[7]，这样提取的有效特征信息能够显著提升视频识别的准确性。目前视频中动作特征主要是利用基于深度学习的方法进行提取，SIMONYAN等[8]提出双流网络模型，分别使用RGB图像和光流图像作为网络输入，提取视频的空间特征和时间特征，最后将2个特征融合来识别视频中的动作。3D卷积网络将视频视为3D时空结构，TRAN等[9]通过开发3D卷积和3D池化层，将2D CNN扩展到3D CNN，并使用3D卷积方法来学习视频中动作的空间和时间特征。

本文提出一种基于注意力机制的3D密集卷积网络模型。使用双流网络基本框架，将原DenseNet网络的二维卷积内核扩展到三维空间，以提升视频序列特征的提取效率。为准确提取视频中人的动作特征，在3D密集卷积网络中加入空间和通道注意力机制，结合时间网络对连续视频序列提取的运动光流进行特征提取，并在双流网络之间对时空网络的相互作用进行建模，最后进行时空特征融合得到视频中人体动作的识别结果。

1 相关工作

视频中动作识别尽管在许多领域中得到了广泛应用，但准确有效的人体动作识别仍然是计算机视觉研究的一个具有挑战的领域[7]。视频中巨大的信息冗余需要大量的存储空间，而且从视频帧中发现具有区分性的信息非常复杂且过程缓慢。近年来，双流网络模型[8]和3D卷积模型[9]在视频中的动作识别上取的了较好的效果。双流卷积网络有一些代表性扩展工作，如ZHANG等[10]使用视频流中的运动矢量而不是光流序列来提高计算速度并实现实时的视频中人体动作识别，FEICHTENHOFER等[11]将空间和时间信息融合的过程从原始的最终分类层更改为网络的中间部分，从而进一步提高了动作识别的准确性，WANG等[12]详细讨论了双流卷积网络的输入、卷积网络结构和训练策略，并提出了时域网（TSN）进一步改善双流卷积网络的结果。另外，LAN[13]和ZHOU[14]的工作也进一步提高了TSN的识别结果。WANG等[15]对双流网络进行跨任务扩展，使用动作识别网络中的信息促进动作预测网络的学习，构建的师生模型在视频中动作的预测任务中取得了较好的效果。

许多研究人员尝试将不同的二维卷积网络扩展到三维时空结构，以学习和识别视频中的人体动作特征。3D卷积网络可以直接从视频中提取时空特征，因此具有较高的识别效率。JI等[16]将卷积神经网络在时间维度上进行了扩充，提出使用3D卷积神经网络进行动作识别。TRAN等[9]提出一种C3D（Convolutional 3D）的标准3D卷积神经网络架构。3D卷积神经网络通过3D内核卷积将多个连续帧堆叠在一起，进而形成多维数据特征来实现对视频动作时空特征的提取。通过3D卷积核这种构造，卷积层中的特征图被连接到前一层中的多个连续帧，从而捕获到运动信息。CARREIRA等[17]将Inception-V1的网络结构从二维扩展到了三维，并提出用于动作识别的双流3D ConvNet。此后出现了一系列基于3D卷积的网络框架[18-20]，ZHU等[21]提出将合并操作从原始空间维扩展到3个时空维，并将双流卷积网络扩展到3D结构。

注意力机制可以利用人类视觉机制进行直观解释，视觉系统中倾向于关注图像中辅助判断的部分信息，并忽略掉不相关的信息[22]，使用注意力机制对于视频中复杂特征信息的提取是有效的。ZHANG等[23]提出空间自注意力机制，使得网络对感兴趣区域有更多的关注。JADERBERG等[24]提出一种注意力模块，通过对图像进行空间变换来提取图像中的关键信息。HU等[25]提出通道注意力模块，通过对不同通道赋予不同的权重来获取每个特征通道的重要程度。WANG等[26]提出一种编解码器注意模块，并将其与ResNet相结合，经过端到端的学习，该网络不仅表现良好，而且对噪声具有较好的鲁棒性。TU等[27]使用时空聚合视频特征进行动作识别，同时设计RGBF模式捕获RGB图像中与动作活动对应的运动突出区域，丢弃视频帧中的冗余特征。WOO等[28]在HU等[25]的基础上提出了卷积块注意力模型（Convolutional Block Attention Module，CBAM）。CBAM结合空间和通道的注意力模块，可以有效提升卷积神经网络的表达能力，在各项任务中都取得了较好的效果。本文为了更好地提取视频中动作的特征，在三维密集卷积神经网络中嵌入改进的CBAM

计算机工程 www.ecaa0.com

注意力模块以提高其性能。本文的主要工作如下：

1）为充分提取视频的时空特征，本文采用双流网络为基本框架，以视频数据的RGB图像和视频帧的光流图像作为输入，分别提取视频动作的时间和空间特征，最后通过融合进行分类识别。

2）为更准确地提取特征并对时空网络之间的相互作用进行建模，在双流网络之间使用跨流连接乘法残差块对时空网络提取的特征进行融合。

3）将DenseNet网络由二维卷积内核扩展到三维时空，加强特征传播与重用，有效地减少了参数的数量。

4）在三维密集卷积网络中加入空间和通道注意力机制，更加准确有效地提取了视频中人的动作特征，在UCF101和HMDB51数据集上取得较好的效果。

2　基于注意力机制的三维密集网络

本文使用密集连接网络作为特征提取网络，引入通道和空间注意力学习机制，提出一种适用于视频中人体动作识别的算法。从特征提取、注意力机制模块、特征融合等3个方面对网络模型进行详细的描述。模型网络框架如图1所示，输入视频帧和相应的光流图片之后分别在空间网络和时间网络中进行特征提取，在提取特征过程中通过注意力模块进一步提升特征提取效率，且时空网络中存在特征交互，最后在分类层进行融合输出最后的识别结果。在特征提取部分，使用双流网络从RGB图像和光流图像两方面对视频帧的表观信息和运动信息进行提取，在双流网络之间用乘法残差块使视频的运动信息与表观信息进行交互，从而更准确地提取视频中的关键特征。为使CBAM注意力模块能够融合到本文网络中，注意力机制模块部分将CBAM模块从二维扩展到了三维，以便更好地嵌入到三维密集网络中。特征融合主要是介绍时空网络之间的交互以及最终分类结果的融合，图1中时间网络到空间网络的连接表示时空网络之间的交互连接。

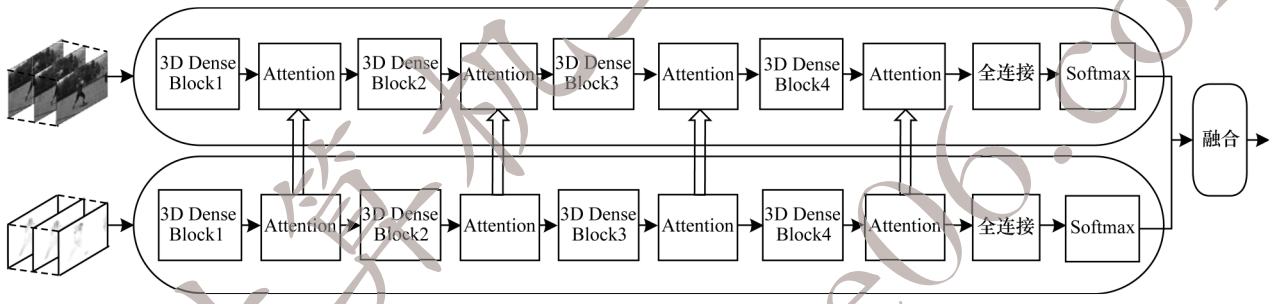


图1　模型网络框架

Fig.1　Model network framework

2.1　三维密集网络

由于DenseNet网络具有参数更少、计算成本低和较好的图像分类性能，本文将传统的DenseNet网络的卷积核由二维扩展到三维卷积核，DenseNet网络使用密集连接方式，每层网络与前层所有网络相连接，连接方式如图2所示。3D卷积核的扩展源于DenseNet和C3D，与T3D中的DenseBlock相似，3D卷积层在空间上对类似于2D卷积的输入特征图进行卷积，并在时间上对连续视频帧之间的时间依赖性进行建模。同样，3D池化层在空间和时间上对输入特征图的大小进行下采样。3D卷积层和3D池化层的内核大小均为$s\times s\times d$，其中，s是空间大小，d是输入视频帧的深度/长度。

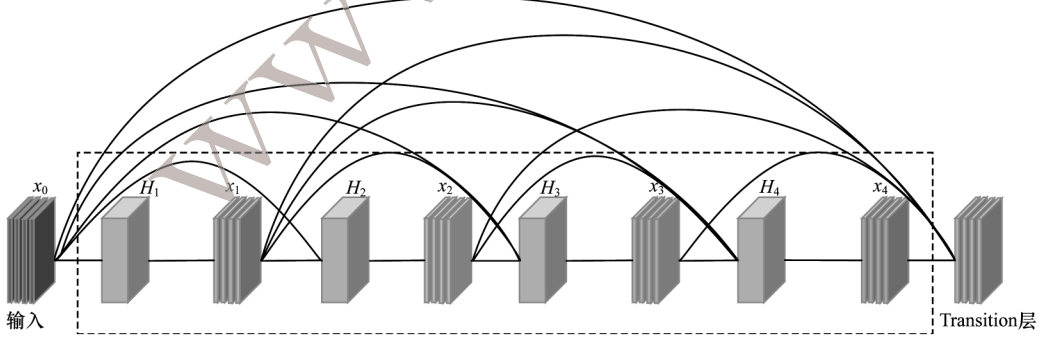


图2　3D DenseNet结构

Fig.2　3D DenseNet structure

3D DenseBlock与二维密集网络中的2D DenseBlock相似，三维密集网络中每一层的特征直接连接到3D DenseBlock中后续各层，连接方式是通过元素级相加，所以，对于一个$l$th层的block，共有

$l(l+1)/2$个连接，同时$l$th层与$(l-1)$th层之间可能实际上包含多个卷积层。$l$th层的输入为$l$th层之前所有3D Dense-Block的输出，$l$th层的输出为x_l，该过程可以描述如下：

$$x_l = H_l[x_0, x_1, \cdots, x_{l-1}] \tag{1}$$

其中：表示前面所有层的输入特征图的密集连接；$H(\cdot)$为非线性转化函数，是BN-ReLU-3DConv组合操作的复合函数。与2D DenseNet相似，3D DenseNet中的每个Dense-Block模块包含批处理归一化(BN)、ReLU激活函数、池化(Pooling)和3D卷积(大小为3×3×3)三部分。与其他网络逐层传递的方式不同，DenseNet网络每层网络与前层所有网络相连接，能够直接获得输入的特征，通过特征在通道上的连接实现特征重用，并且鲁棒性更好。

2.2 注意力机制

虽然三维卷积神经网络在动作识别领域已有大量的研究，但仍然是一个具有挑战性的问题。对于视频来讲，信息主要包含在视频序列关键帧的关键区域中。但在实际实验中，3D卷积神经网络往往缺乏提取这类关键信息的能力[23-24]。研究结果表明，注意力机制是一种忽略不重要信息而提取关键信息的方法，且在图像领域，许多网络模型都引入了注意力机制，并取得了良好的效果[25-26]。因此，本文提出在三维卷积神经网络中引入注意机力机制，以便更好地提取视频中关键信息的特征。

本节描述了双流网络中的三维CBAM模块。基于传统的CBAM[26]将其中的卷积池化操作从二维扩展到了三维，生成一个三维CBAM的注意力模块，结合3D DenseNet网络能够有效提升对视频帧特征的提取，三维CBAM模块可以对特定区域选择性聚焦来更好地捕获视频帧中的特征。注意力模块框架如图3所示，注意力模块由通道注意力模块与空间注意力模块串联组成，在深度网络的每个卷积块上，使用注意力模块能够有效地细化网络中的特征映射，提高特征提取效率。

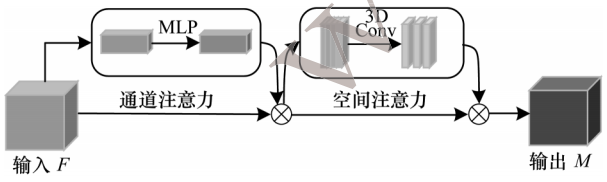


图3 三维卷积块注意力模块

Fig.3 3D convolutional block attention module

通道注意力模块：与传统CBAM模块类似，通道注意力模块主要是确定视频帧中关键区域“是什么”，输入的特征图为$F \in \mathbb{R}^{T\times C\times H\times W}$，如图4所示，采用3D最大池化和3D平均池化操作相结合的方法压缩输入特征映射的空间维数，提高了通道注意力的效率。通过池化操作后生成2个不同的特征描述符$F^C_{3\text{Davg}}$、$F^C_{3\text{Dmax}}$，输入到含有一个隐藏层的多层感知器(MLP)中生成通道注意力特征图$M_C(F)$，经过通道注意力机制模块中激活函数σ后得到各通道的权重系数，这样对不同的通道特征图分配不同的权值，对包含有效关键信息的通道分配较高的权值，对其他通道分配较低的权值，从而确定了视频帧中包含有效特征的通道。为了减少参数的开销，将隐藏层的特征维数为C/r(r为约简率)，每个特征描述符通过多层感知器后，使用按元素进行求和的方法合并输出特征向量，通道注意力的计算公式如下：

$$M_C(F) = \sigma(W_1(W_0(F^C_{3\text{Davg}})) + W_3(W_2(F^C_{3\text{Dmax}}))) \tag{2}$$

其中：σ为sigmoid激活函数；$W_0, W_2 \in \mathbb{R}^{C/r\times C}$与$W_1, W_3 \in \mathbb{R}^{C\times C/r}$中的$W_0$、$W_1$和$W_2$、$W_3$是MLP的权重，对于2个输入，分别属于不同的多层感知器(MLP)。最后将输入的特征图与通道注意力特征图进行元素的加和操作，从而对视频帧中关键通道进行有效提取。

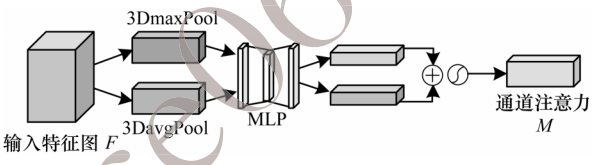


图4 通道注意力模块

Fig.4 Channel attention module

空间注意力模块：与通道注意力机制不同，空间注意更多关注图像中的有效特征“在哪里”，这是对通道注意提取的特征的补充。空间注意力机制是对视频帧中不同的空间区域分配不同的权重，对视频帧中存在动作的关键区域分配较高的权重，从而确定关键信息在视频帧中的位置。空间注意力模块如图5所示，本文使用3D池化内核对空间注意力模块进行扩展，并使用最大池化和平均池化操作得到$F^S_{3\text{Dmax}}$和$F^S_{3\text{Davg}}$，将其连接得到一个有效的特征描述符$[F^S_{3\text{Dmax}}; F^S_{3\text{Davg}}]$，经过一个三维卷积层生成一个空间注意力特征图$M_S(F)$，空间注意力的计算过程如下：

$$M_S(F) = \sigma\left(f^{7\times 7\times 7}\left[F^T_{3\text{Davg}}; F^T_{3\text{Dmax}}\right]\right) \tag{3}$$

其中：$f^{7\times 7\times 7}$为三维卷积运算，其3D卷积过滤器大小为$7\times 7\times 7$。在空间注意力模块中对输入特征图与空间注意力特征图进行累积合并操作，给视频帧中的关键区域较高的权重，从而提取出视频帧中的关键部分。

计算机工程 www.ecice06.com

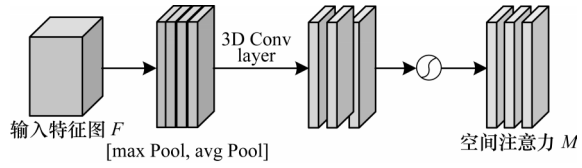
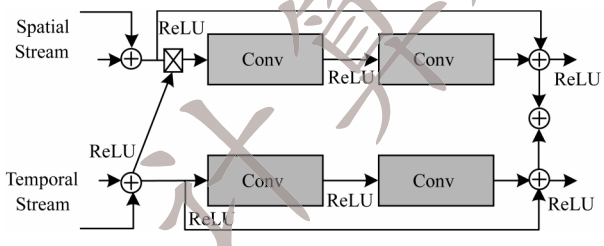


图5　空间注意力模块

Fig.5　Spatial attention module

2.3　特征融合

在本文的网络模型中，对于输入的视频，通过对视频均匀采样RGB帧和对应的光流帧，分别输入到具有相同CNN结构的时空网络，采用双流结构能够更好地模拟视频中运动的外观和运动特征的相关性。传统的双流网络是时空网络提取相应的特征后进行Softmax预测，最后计算最终的分类结果，由于这种结构并行地提取了视频的外观特征和运动特征，在最终融合之前没有交互作用，无法对视频帧中细微的时空线索进行建模。因此，在时间网络与空间网络之间加入跨流连接来进一步对视频帧中细微的时空线索进行建模，跨流连接使用乘法残差块对时空网络特征进行融合，跨流连接模块方式如图6所示。



6　跨流连接方式

Fig.6　Cross-stream connection mode

在提取特征部分对网络进行从空间网络到时间网络的跨流连接特征融合，分别进行4次融合，每次融合都在池化操作之后，融合特征后通过注意力机制模块准确有效地提取运动特征，融合方式如式(4)所示。最后在分类层对时空网络进行融合，时空网络分别输出一个Softmax层，最后通过加权融合将两个Softmax层的输出结果进行融合，得到视频中人体动作识别的最终识别准确率。同时，在反向传播过程中，来自每一个网络的信号可以共同参与并影响梯度，这有助于缓解原双流网络在建模细微时空相关性方面的不足。

$$x_{l+1}^t = f(x_l^t) + F(x_l^t \odot f(x_l^s), W_l) \tag{4}$$

其中：$\odot$表示按元素进行相乘的符号；x_l^t和x_l^s分别表示时间网络和空间网络第l层的输入；x_{l+1}^t表示时间网络第$l+1$层的输入；$f(x)$代表ReLU激活函数；$F(x)$是代表卷积层和ReLU激活层的运算符。

3　实验结果与分析

3.1　数据集

数据集主要包括以下2种：

1）UCF101[29]数据集。UCF101数据集是一个人体运动视频数据，包含101类动作，共有13 320个视频段，101个动作类别中的视频分为25组，每组可包含4~7个动作视频，主要分为人与物体互动、人体动作、人与人互动、乐器演奏、体育运动等5种类别，如化妆、打字、吹头发、骑马、跳高等动作。它在动作方面提供了最大的多样性，并且在相机运动、物体外观和姿势、物体比例、视点、杂乱的背景、照明条件等方面存在较大的变化。部分动作示意图如图7所示。



图7　UCF101数据集部分动作示意图

Fig.7　Schematic diagram of partial actions in UCF101 dataset

2）HMDB51[30]数据集。包含51类动作，共有6 849个视频段，分辨率为320像素×240像素。主要分为面部动作、面部操作、身体动作、交互动作、人体动作等5类，如微笑、吸烟、拍手、梳头、打球、拥抱等动作。该数据集的视频大多来源于电影剪辑片段，小部分来自YouTube等视频网站，像素相对较低。部分动作示意图如图8所示。



图8　HMDB51数据集部分动作示意图

Fig.8　Schematic diagram of partial actions in HMDB51 dataset

3.2　实验过程

本文将数据集的最终识别准确率值作为人体动作识别模型的评估指标。对于UCF101和HMDB51这2个数据集，采用原始训练/测试拆分，并遵循这2个数据集提供的标准评估协议，将3个拆分的平均值作为最终结果，每种拆分方式都是将数据集分为70%的训练集和30%的测试集。

计算机工程 www.ecice06.com

对于双流网络，使用文献[11]中预先计算的视频RGB和光流作为模型的输入，输入视频后对视频帧进行均匀取样，以保证视频中动作的连续性，取样视频的RGB图像和相对应的光流帧用作模型的输入。对于空间网络，采用数据库ImageNet[31]对其进行预训练。对于时间网络，采用TL-V1[32]方法提取RGB数据的光流运动信息，通过线性变换将光流数据离散到[0,255]上，以便与RGB数据同区间。

由于人体动作数据集中取样的视频帧有限，在较深的网络结构中容易产生过拟合现象，因此对输入的视频帧采用与文献[33]相同的数据增强策略，对视频帧进行随机位置裁剪、适当翻转等操作扩大训练集。视频的最大裁剪大小（即图像的高度和宽度）设置为224×224，以减少GPU内存的使用。在Linux系统搭建Ubuntu16.04的TensorFlow平台进行实验。由于深度神经网络容易陷入过拟合现象，因此本文将模型中空间网络和时间网络dropout层的丢失率分别设置为0.7和0.8，设置初始学习率为10^{-3}，并在之后的训练中对学习率进行适当调整。在对比实验中，除了对该模型在数据集上的识别准确率进行分析外，还对是否添加注意力机制及跨流连接方式进行了消融实验。

3.3 结果分析

本文对基于注意力机制的三维密集网络模型的注意力机制部分进行了可视化，将通过注意力机制生成的热图与原始图像进行结合，可以看出，通过注意力机制模块后，能够准确地识别出人体动作的关键特征，观察可视化图片可以发现，人体动作的关键信息主要集中在运动的部分。可视化结果如图9所示，红色部分表示学习到的注意力机制的预测，代表图像中具有关键运动信息的部分，蓝色部分代表视频中的背景部分（彩色效果见《计算机工程》官网HTML版）。表1为是否添加注意力模块对最终识别准确率的影响。从表1可以看出，在时间网络和空间网络上使用添加注意力机制的网络模型比不使用注意力机制的动作识别准确率高，表明添加注意力机制的网络模型比视频中人体动作识别任务更有效。

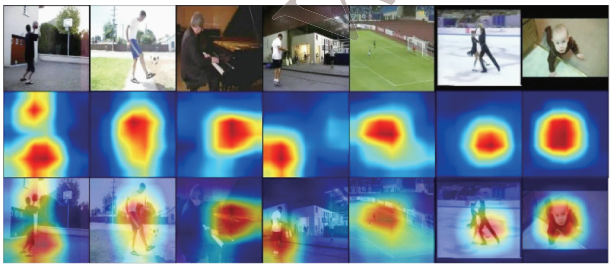


图9 注意力机制可视化图像

Fig.9 Attention mechanism visualization images

表1 添加注意力模块对最终识别准确率的影响

Table 1 Effect of adding attention module on the final recognition accuracy %

添加注意力机制	UCF101	HMDB51
否	92.73	68.87
是	94.52	69.64

表2为本文方法模型和其他经典的动作识别方法模型在UCF101和HMDB51数据集识别准确率的结果对比。文献[8-9]分别提出了双流网络模型和C3D卷积网络模型，文献[11]提出的时空融合网络，其网络结构是双流VGG模型，分析了时空网络的不同融合方式，为本文提供了工作基础。文献[34]引入一种新的时域层为可变时域卷积核深度建模。实验结果表明，本文模型对视频中人体动作的识别有较高的准确率。

表2 不同模型在UCF101数据集和HMDB51数据集上的识别准确率对比

Table 2 Comparison of the recognition accuracy of different models on UCF101 dataset and HMDB51 dataset %

模型	UCF101	HMDB51
Two-Stream[8]	88.0	59.4
C3D[9]	88.0	59.4
Fusion[11]	92.5	65.4
ResNet3D	86.1	55.6
DenseNet3D	88.9	57.8
T3D[34]	93.2	63.5
本文模型	94.5	69.6

表3为3D DenseNet与C3D两种模型在UCF101和HMDB51数据集上识别准确率的结果对比。通过实验可以看出，基于C3D与DenseNet网络改进的3D DenseNet网络在时间网络、空间网络以及双流网络中都取得了较好的效果。实验中对两种3D网络模型进行对比，由于未使用注意力机制，避免了注意力机制对该实验的影响。从表3还可以看出，与空间网络提取的外观信息相比，时间网络提取的光流运动信息识别准确率更高。这也说明了在视频中的人体动作识别任务中，运动信息更加重要。

表3 3D DenseNet与C3D两种模型在UCF101和HMDB51数据集上识别准确率的结果对比

Table 3 Comparison of results of classification accuracy between 3D DenseNet and C3D models on UCF101 and HMDB51 datasets %

图像	模型	UCF101	HMDB51
RGB	C3D	81.5	46.2
	3D DenseNet	83.6	47.3
Flow	C3D	85.9	53.3
	3D DenseNet	85.8	55.3
Two-stream	C3D	89.3	61.6
	3D DenseNet	92.7	68.9

计算机工程 www.ecice06.com

从表3可以看出，结合光流运动信息的双流网络能够有效改善单支网络的识别准确率。本文对时空网络在最终分类层的融合进行探究，时空网络融合的识别准确率如表4所示，设置不同时间网络和空间网络分类的识别置信度，以便对时空网络的识别准确率进行线性加权融合得到最终的识别结果，在UCF101数据集的3个拆分（split）方式下进行实验，由表4可以看出，时空网络的识别置信度设置为1∶1时，最终的融合识别准确率优于1∶1.2和1∶1.5的情况。

表4　UCF101数据集上时空网络融合的识别准确率

Table 4　The recognition accuracy of spatiotemporal network fusion on the UCF101 dataset　%

拆分方式	UCF101数据集		
	置信度为1:1	置信度为1:1.2	置信度为1:1.5
split1	94.43	93.87	94.03
split2	94.67	94.33	94.26
split3	94.46	94.40	94.17
平均	94.52	94.20	94.15

本文在特征融合的跨流连接如图6所示，采用乘法残差块对时空网络特征进行融合，本文将加法与乘法两种融合方式进行比较，对比结果如表5所示。使用加法进行融合，在UCF101和HMDB51的第1个分段上分别产生9.73%和43.86%的测试误差，相比之下使用乘法方式进行融合，测试误差降低到了8.56%和39.43%。因此，在进行跨流连接的选择上本文采用的是乘法残差块进行特征融合。

表5　不同跨流连接方式在UCF101和HMDB51数据集上的测试误差率

Table 5　Different cross-stream connection modes test error rate on UCF101 and HMDB51 datasets　%

连接方式	UCF101	HMDB51
加法融合	9.73	43.86
乘法融合	8.56	39.43

表6为双流网络中对不同跨流连接方向的研究，使用乘法残差块进行特征融合，分别对空间网络到时间网络、时间网络到空间网络以及时空网络相互连接3种连接方向进行了实验。实验结果发现，将跨流连接的方向设置为从时间网络到空间网络能够有效减少测试误差，但也发现，从空间网络的动作表观信息流到时间网络的光流信息进行连接容易发生表观信息的严重过拟合。

表6　在UCF101和HMDB51数据集上不同跨流连接方向的测试误差率对比

Table 6　Comparison of test error rates of different cross-stream connection directions on UCF101 and HMDB51 datasets　%

连接方向	UCF101	HMDB51
时间←空间	18.69	50.27
时间→空间	8.56	39.43
时间↔空间	16.31	49.23

4　结束语

本文提出一种基于注意力机制的三维DenseNet双流网络结构用于视频中的人体动作识别。该模型针对光流图片中动作的运动信息，有效提取视频序列的时间特征，结合视频动作的表观信息，通过适当的卷积层特征融合和分类层融合，提高视频中人体动作的识别准确率。同时，在网络中加入注意力机制和时空网络的跨流连接，通过提取动作关键特征对时空线索进行建模，进一步提升特征提取效率，从而提高识别准确率。实验结果表明，本文模型能够在UCF101数据集和HMDB51数据集上取得较好的识别准确率，并且能够减少模型的参数，加快训练测试过程，具有更好的泛化能力。本文模型由于训练了时间和空间两个网络，导致降低模型参数的效果有限，下一步将对时空网络进行更好的融合，减少模型参数，从而提高识别准确率。

参考文献

[1] BEN M A，ZAGROUBA E. Abnormal behavior recognition for intelligent video surveillance systems：a review[J]. Expert Systems with Applications，2018，91(1)：480-491.

[2] WANG P C，LI W Q，OGUNBONA P，et al. RGB-D-based human motion recognition with deep learning：a survey[J]. Computer Vision and Image Understanding，2018，9(1)：1-22.

[3] PRESTI L L，CASCIA M L. 3D skeleton-based human action classification：a survey[J]. Pattern Recognition，2015，53：130-147.

[4] WANG L，HUYNH D Q，KONIUSZ P. A comparative review of recent kinect-based action recognition algorithms[J]. IEEE Transactions on Image Processing，2020，29(3)：15-28.

[5] SAIF S，TEHSEEN S，KAUSAR S. A survey of the techniques for the identification and classification of human actions from visual data[J]. Sensors，2018，18(11)：39-49.

[6] MONTES A，SALVADOR A，PASCUAL S，et al. Temporal activity detection in untrimmed videos with recurrent neural networks[EB/OL]. [2020-08-10]. https://arxiv. org/pdf/1608. 08128. pdf.

[7] 张瑞，李其申，储珺. 基于3D卷积神经网络的人体动作识别算法[J]. 计算机工程，2019，45(1)：259-263.

计算机工程 www.ecice06.com

ZHANG R, LI Q S, CHU J. Human action recognition algorithm based on 3D convolution neural network[J]. Computer Engineering, 2019, 45(1): 259-263. (in Chinese)
[8] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[EB/OL]. [2020-08-10]. https://arxiv.org/abs/1406.2199.
[9] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2015: 4489-4497.
[10] ZHANG B, WANG L, WANG Z, et al. Real-time action recognition with deeply transferred motion vector CNNs[C]//Proceedings of 2016 IEEE International Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2018: 1-10.
[11] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2016: 1933-1941.
[12] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition [C]//Proceedings of IEEE ECCV'16. Washington D.C., USA: IEEE Press, 2016: 20-36.
[13] LAN Z, ZHU Y, HAUPTMANN A G. Deep local video feature for action recognition[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2017: 1219-1225.
[14] ZHOU B, ANDONIAN A, TORRALBA A. Temporal relational reasoning in videos[EB/OL]. [2020-08-10]. https://arxiv.org/pdf/1711.08496.pdf.
[15] WANG X, HU J F, LAI J H, et al. Progressive teacher-student learning for early action prediction[C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2019: 235-247.
[16] JI S, YANG M, YU K. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.
[17] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE Press, 2017: 4724-4733.
[18] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]// Proceedings of 2017 IEEE International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2017: 367-381.
[19] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2018: 365-378.
[20] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks [C]//Proceedings of IEEE CVPR'17. Washington D.C., USA: IEEE Press, 2017: 443-456.
[21] ZHU J, ZOU W, ZHU Z. End-to-end video-level representation learning for action recognition [C]// Proceedings of the 24th International Conference on Pattern Recognition. Washington D.C., USA: IEEE Press, 2017: 234-245.
[22] CHAUDHARI S, POLATKAN G, RAMANATH R, et al. An attentive survey of attention models[EB/OL]. [2020-08-10]. https://arxiv.org/abs/1904.02874v3.
[23] ZHANG J, XIE Y, XIA Y, et al. Attention residual learning for skin lesion classification[J]. IEEE Transactions on Medical Imaging, 2019, 38(9): 2092-2103.
[24] MAX J, SIMONYAN K, ZISSERMAN A. Spatial transformer networks[C]//Proceedings of ECCV'15. Berlin, Germany: Springer, 2015: 2017-2025.
[25] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.
[26] WANG F, JIANG M, QIAN C, et al. Residual attention network for image classification[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Press, 2017: 189-197.
[27] TU Z G, LI H Y, ZHANG D J, et al. Action-stage emphasized spatio-temporal VLAD for video action recognition[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2799-2812.
[28] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of International Conference on Computer Vision. Washington D.C., USA: IEEE Press, 2018: 3-19.
[29] SOOMRO K, ZAMIR A R, SHAH M. Ucf101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2020-08-10]. https://arxiv.org/pdf/1212.0402.pdf.
[30] KUEHNE H, JHUANG H, STIEFELHAGEN R, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of International Conference on High Performance Computing in Science and Engineering. Berlin, Germany: Springer, 2013: 571-582.
[31] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D.C., USA: IEEE Computer Society, 2009: 248-255.
[32] PÉREZ J S, MEINHARDT-LLOPIS E, FACCIOLO G. TV-L1 optical flow estimation[J]. Image Processing on Line, 2013(3): 137-150.
[33] ZINKEVICH M, WEIMER M, SMOLA A J, et al. Parallelized stochastic gradient descent[C]//Proceedings of IEEE DBLP'11. Washington D.C., USA: IEEE Press, 2011: 485-496.
[34] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3D ConvNets: new architecture and transfer learning for video classification[EB/OL]. [2020-08-10]. https://arxiv.org/pdf/1711.08200.pdf.

编辑 索书志

计算机工程 www.ecice06.com