融合双流三维卷积和注意力机制的动态手势识别

王粉花*①②③ 张 强① 黄 超① 张 苒①

①(北京科技大学自动化学院 北京 100083)

②(北京科技大学人工智能研究院 北京 100083)

③(北京市工业波谱成像工程中心 北京 100083)

摘 要：得益于计算机硬件以及计算能力的进步，自然、简单的动态手势识别在人机交互方面备受关注。针对人机交互中对动态手势识别准确率的要求，该文提出一种融合双流3维卷积神经网络(I3D)和注意力机制(CBAM)的动态手势识别方法CBAM-I3D。并且改进了I3D网络模型的相关参数和结构，为了提高模型的收敛速度和稳定性，使用了批量归一化(BN)技术优化网络，使优化后网络的训练时间缩短。同时与多种双流3D卷积方法在开源中国手语数据集(CSL)上进行了实验对比，实验结果表明，该文所提方法能很好地识别动态手势，识别率达到了90.76%，高于其他动态手势识别方法，验证了所提方法的有效性和可行性。

关键词：动态手势识别；深度学习；双流3维卷积神经网络；注意力机制；BN层

中图分类号：TP183 **文献标识码**：A **文章编号**：1009-5896(2021)05-1389-08

DOI: 10.11999/JEIT200065

Dynamic Gesture Recognition Combining Two-stream 3D Convolution with Attention Mechanisms

WANG Fenhua①②③ ZHANG Qiang① HUANG Chao① ZHANG Ran①

①(*School of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*)

②(*Institute of Artificial Intelligence, University of Science and Technology Beijing, Beijing* 100083, *China*)

③(*Beijing Engineering Research Center of Industrial Spectrum Imaging, Beijing* 100083, *China*)

Abstract: Benefits from the progress of computer hardware and computing power, natural and simple dynamic gesture recognition gets a lot of attention in human-computer interaction. In view of the requirement of the accuracy of dynamic gesture recognition in human-computer interaction, a method of dynamic gesture recognition that combines Two-stream Inflated 3D (I3D) Convolution Neural Network (CNN) with the Convolutional Block Attention Module (CBAM-I3D) is proposed. In addition, relevant parameters and structures of the I3D network model are improved. In order to improve the convergence speed and stability of the model, the Batch Normalization (BN) technology is used to optimize the network, which shortens the training time of the optimized network. At the same time, experimental comparisons with various Two-stream 3D convolution methods on the open source Chinese Sign Language (CSL) recognition dataset are performed. The experimental results show that the proposed method can recognize dynamic gestures well, and the recognition rate reaches 90.76%, which is higher than other dynamic gesture recognition methods. The validity and feasibility of the proposed method are verified.

Key words: Dynamic gesture recognition; Deep learning; Two-stream 3D Convolution Neural Network (CNN); Attention mechanism; Batch Normalization (BN) layer

收稿日期：2020-01-16；改回日期：2020-12-06；网络出版：2020-12-18

*通信作者： 王粉花 wangfenhua@ustb.edu.cn

基金项目：国家重点研发计划重点专项(2017YFB1400101-01)，北京科技大学中央高校基本科研业务费专项资金(FRF-BD-19-002A)

Foundation Items: The National Key Research and Development Project of China (2017YFB1400101-01), The Fundamental Research Funds for the Central Universities (FRF-BD-19-002A)

1 引言

随着深度学习以及虚拟现实技术的发展，人机交互技术正逐渐从以机器和计算机为中心转移到以人为中心[1]。手势是人类最常用的人机交互方式(Human -Computer Interaction, HCI)之一，也是正常人与聋哑人以及聋哑人之间沟通交流的重要方式[2]。手势识别是人机交互领域的一个前沿课题和研究热点。手势本身具有灵活性、不固定性和多异性，所以手势识别是一个富有挑战性的多学科交叉的研究方向。手势识别根据数据集是图片还是视频可分为静态手势识别和动态手势识别，静态手势识别只关注某个时间点的外形特征，动态手势识别则需关注一段时间的动作，增加了时间信息和动作特征。相比较于静态手势识别，动态手势识别更贴近人的表达习惯、更加具有现实意义。

早期的手语识别主要依靠可穿戴设备和人工经验提取特征，但是可穿戴设备需要人们佩戴特定的设备，影响了人们的活动自由，具有一定的局限性[3]。Parcheta等人[4]使用基于隐马尔可夫模型(Hsidden Markov Model, HMM)的手势识别，在包含91个手语单词数据集上取得了比较高的准确率。运用HMM进行手语识别有两种情况：一是为每个单词构造一个HMM模型[5]，二是为多个单词建立一个HMM模型[6]。将手语视频中的手形和运动轨迹作为信息提取特征来计算手语识别。但是人工提取特征非常耗时耗力，并且需要领域内的专家才能设计出好的分类特征。随着卷积神经网络的不断发展，利用深度学习实现基于视频的动态手势识别受到了研究者的关注，视频处理要求考虑时域连接。在这方面，循环神经网络(Recurrent Neural Network, RNN)显示出其巨大的优势。然而RNN在特征提取方面并不令人满意，为此一些研究者选择先用卷积神经网络(Convolution Neural Network, CNN)来提取图像特征，生成特征向量并输入RNN进行计算[7]。在CNN部分，可以使用预先训练好的CNN网络进行图像特征提取[8]，这大大提高了网络的训练效率。考虑到视频的时空信息，3维卷积神经网络(3D-CNN)应运而生。如Tran等人[9]提出的C3D网络，使用了3D-CNN代替了传统卷积神经网络中的2D-CNN，将时间信息也考虑在内。Chen等人[10]提出了MFNet。还有参考ResNet等2D-CNN网络产生的新3D-CNN结构，这些网络在行为识别上取得了很好的成绩，同样也可以将它们应用于手语识别[11]。在动作识别和手语识别之间有许多相似之处，3D-CNN为手语识别提供了一种很好的方法[12]。然而仅输入原始图像不足以获得足够多的特征。许多研究者开始利用卷积神经网络和多流输入[13]来获得更多的信息，如光流[14]信息、深度信息、后信息[15]等，这有利于提高网络识别的精度。借鉴双流网络和3D-CNN的优点，Deep mind团队提出了I3D网络[16]，扩展2D卷积Inception v1为3D 结构，同时将光流单独作为一个分支和原始图像形成双流网络，RGB图像和光流分开训练。I3D是目前行为识别最好的网络之一。

本文提出了一种在I3D网络的基础上融合注意力机制CBAM[17]的动态手势的识别方法。在CSL[18]数据集上进行实验，首先从RGB视频中提取光流特征，然后将RGB视频、光流信息分解成帧图片集合。将数据按照3:1:1随机划分为训练集、验证集和测试集，最后将这两种数据送入同一个网络结构中进行训练。本文所提出方法的识别准确率达到了90.76%，实现了对动态手势很好的识别。

2 网络结构

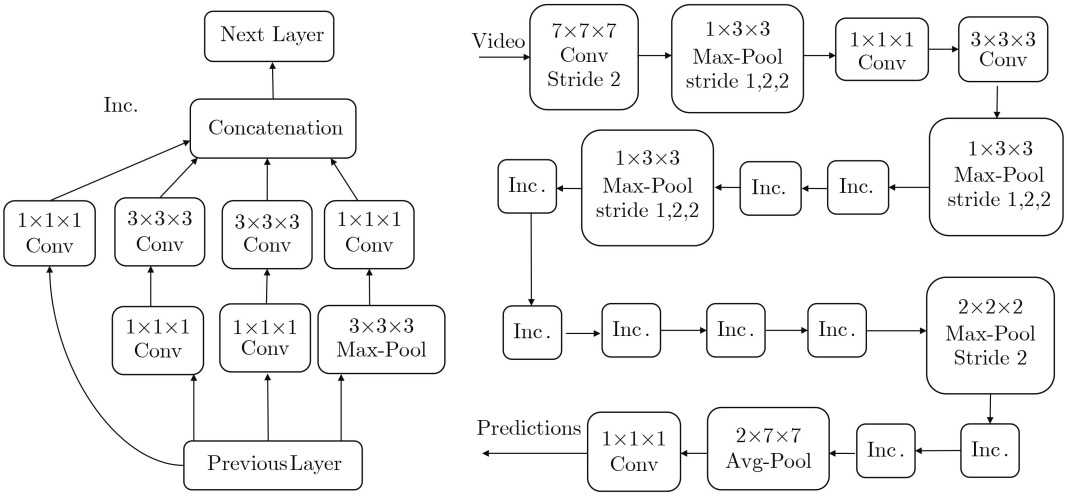
2.1 I3D网络

I3D网络是由Deep mind团队提出的最新3D卷积网络之一，在总结以往主流视频动作识别模型优缺点的基础上，把双流的思想加入3D卷积当中，其中一个3D卷积网络接收RGB信息，另一个3D卷积网络接收优化后的平滑光流信息。3D卷积结构由2D卷积Inception v1扩展而来，如图1(a)所示，具体做法是把2D模型中的卷积核参数在时间维度上不断复制，形成3D卷积核的参数，之后除以N，保证网络输出和2D卷积上一样，而其他的非线性层结构都与原来的2D卷积模型一样，卷积核和池化均增加了时间维度。为了使时间维度不会因过快或过慢的缩减导致捕捉不到动态场景的发生，在前两个池化层上将时间维度的步长设为1，空间维度是2×2，最后池化层的维度是2×7×7。网络的具体连接方式如图1(b)所示。虽然3D卷积可以直接学习视频的时间特征，但是它只执行纯粹的前向传播，而光流算法在一定程度上提供了迭代的思想在里面，将光流加进来后可以提高网络的识别精度。

2.2 CBAM网络

近年来很多论文验证了在神经网络中引入注意力机制能够提高网络模型的特征表达能力。注意力不仅能告诉网络模型该注意什么，而且也能增强图像特定区域的表征能力。由Woo等人[17]提出的CBAM是一种结合了空间(spatial)和通道(channel)的注意力机制模块，如图2所示。相较于SENet[19]只关注通道的注意力，CBAM显得更加全面。

通道注意力也就是特征注意力，而空间注意力则是反映输入结果在空间维度上的重要程度。两个模块结合，就对输入进行了立体化的注意力处理，



(a) 3D Inception v1　　(b) I3D网络的连接方式

图 1 I3D网络

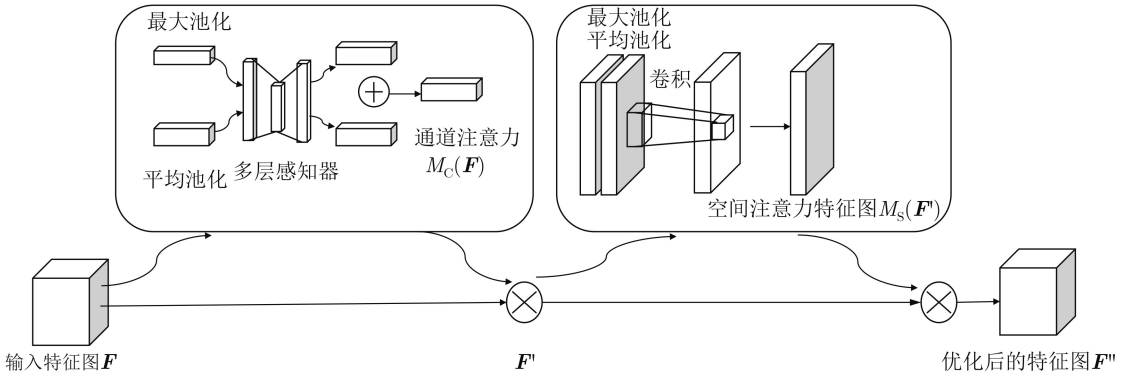


图 2 注意力机制CBAM模型

相当于一个筛子，在输入神经网络之前，先将输入进行一遍筛选，选出最重要的特征。

$$M_{\mathrm{C}}(\boldsymbol{F})=\sigma\{\mathrm{MLP}[\mathrm{AvgPool}(\boldsymbol{F})]+\mathrm{MLP}[\mathrm{MaxPool}(\boldsymbol{F})]\} \tag{1}$$

其中，$M_{\mathrm{C}}(\boldsymbol{F})$为通道注意力模块，$\sigma$为Sigmod激活函数，MLP()为多层感知器，AvgPool()和MaxPool()分别为平均值池化和最大值池化。这个模块相当于一个滤波器，重要的通道权重较大，不重要的通道权重较小，就实现了在特征维度上的注意力机制。

空间注意力模型关注的是有用的信息“在哪里”。首先应用平均值池化和最大值池化对输入进行处理，不过这次是在通道维度上进行压缩采样，生成两个2维的空间矩阵，然后把它们叠在一起，输入一个卷积层，进行权重的学习优化，这就又生成了一个空间注意力滤波器，如式(2)所示

$$M_{\mathrm{S}}(\boldsymbol{F'})=\sigma\{\mathrm{f}\{[\mathrm{AvgPool}(\boldsymbol{F'});\mathrm{MaxPool}(\boldsymbol{F'})]\}\} \tag{2}$$

其中，$M_{\mathrm{S}}(\boldsymbol{F'})$为空间注意力模块，f为卷积层运算。输入特征图首先与通道注意力模块点乘，再与空间注意力模块点乘，得到最后经过CBAM注意力处理后的特征图，如式(3)和式(4)所示

$$\boldsymbol{F'}=M_{\mathrm{C}}(\boldsymbol{F})\cdot\boldsymbol{F} \tag{3}$$

$$\boldsymbol{F''}=M_{\mathrm{S}}(\boldsymbol{F'})\cdot\boldsymbol{F'} \tag{4}$$

其中，$\boldsymbol{F}$为输入特征图，$\boldsymbol{F''}$为输出特征图。

2.3 BN层

深层神经网络的训练很复杂，在训练过程中每一个隐藏层的参数改变都会影响后一层的输入，导致每一批次的数据分布也随之发生改变，使得神经网络需要在每次迭代中学习不同的数据，大大增加了网络学习的难度以及网络过拟合的风险。为了解决以上问题，由Ioffe等人[20]在2015年提出批量归一化处理方法。通常网络的训练采用mini-batch训练法，将整个数据集划分为多个批次，每个批次包含很多组数据，训练的时候以批次为单位进行数据的输入，并进行1次优化。这样做的好处是，每次迭代整个数据集，可以进行多次优化，而不是1次迭代只优化1次，加快了网络训练的速度；而且每次以批次为单位进行运算，引入两个额外参数来实现批量归一化操作。其计算公式为

$$\mu_\beta=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{5}$$

$$\sigma_\beta^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_\beta)^2 \tag{6}$$

$$\widehat{x}_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \varepsilon}} \tag{7}$$

$$y_i = \gamma\widehat{x}_i + \beta \tag{8}$$

其中，式(5)计算输入数据的均值，式(6)计算输入数据的方差，式(7)进行数据的标准化，式(8)进行数据的偏移：数据经过BN层的处理，会更加靠近原点分布，使得激活函数能够获得较大梯度；同时数据也由较为稀疏的分布变得更加紧密，利于神经网络去进行分类。因为密集、整齐的数据，往往更容易被拟合，且不容易出现过拟合的情况(神经网络学习到了数据分布中混乱的不必要的部分)。各种不同来源不同分布的数据，经过BN层的标准化处理，分布会更趋于统一，也有利于提高网络的泛化性能，因此BN层在提高网络训练效率的同时，还可以提升网络的性能，如今被越来越多的研究者所采用。

2.4 CBAM-I3D网络

本文提出的CBAM-I3D网络的最小结构如图3(a)所示，是在I3D的3D Inception v1模块的Concatenation层后加入了CBAM。通过这种融合网络既可以实现原始输入信息的无损传输，又可以自动学习得到图像的空间位置和通道的重要程度，然后根据重要程度去增强有用特征并抑制无用特征，从而实现空间和通道的自适应校准，添加CBAM注意力机制对网络结构的影响不大，却能使网络学习到图像中更加重要的通道特征和空间位置。其中每一个卷积运算单元都是由3D卷积、BN层和ReLU激活函数层组成的。BN层和ReLU函数缓解了梯度消失的问题，也减轻了深层神经网络的退化，使这个小单元可以多次重复堆叠使用，构建一个足够深的神经网络。由于小的卷积核可以获得更好的效果，故这个结构单元中的所有卷积核尺寸都没有超过3。除了加入注意力机制CBAM外，本文还对I3D的网络结构做了一些修改：(1)去掉前两个最大池化层，防止由池化操作导致图像的低级特征丢失。(2)去掉最后的平均池化操作，只保留$1\times1\times1$卷积，这样是为了在减少大量参数的情况下保留图像的全局信息，增加了网络的鲁棒性。网络具体的连接方式如图3(b)所示，网络权重采用标准正态分布(均值为0，方差为1)随机初始化。

2.5 双流CBAM-I3D网络

本文提出的双流CBAM-I3D网络的两个输入分别选择原始的RGB图像和光流图像，整体的网络结构如图4所示。

使用双流网络的优势在于可以综合考虑原始图像中丰富的空间特征信息和光流图像中的运动特征信息，提取更多的特征用来分类，增加识别精度。

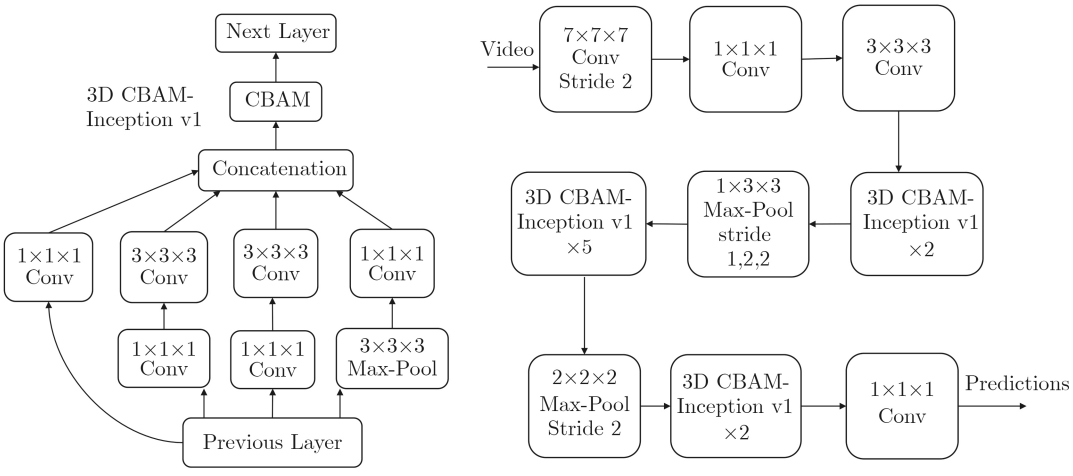
3 实验分析

3.1 运行环境介绍

本次实验的硬件配置为Intel Xeon E5-2660 v4 CPU, 64 GB内存，GPU显卡为2组NVIDIA TITAN Xp, 24 GB显存。软件环境为64位Ubuntu 16.04操作系统，CUDA 8.0.61, cuDNN 5.1.10，深度学习框架为PyTorch ，版本为1.0.1, Python版本为3.5.3。

3.2 数据集

实验使用的手语数据集CSL包括一个手语单词数据集、一个手语连续词句数据集。这两个数据集都包括RGB视频和深度图像视频。本文使用的主



(a) 3D CBAM-Inception v1　　(b) CBAM-I3D网络的连接方式

图 3 CBAM-I3D网络

要是手语单词数据集的RGB视频，共有100类手语单词，5013个RGB视频。为了防止过拟合，这里将它按照3:1:1分成了训练集、验证集和测试集两部分。为了组建双流网络，本文还制作了一个光流图的数据集，用于作为第2部分网络的输入，将在下面做详细介绍。

3.3 数据预处理

3.3.1 RGB视频预处理

CSL动态手势数据集原始视频长度约为2 s，帧宽度为1280，帧高度为720，帧速率为30 帧/s。本文使用的神经网络是以I3D为基础的，需要输入连续的图像帧。所以先将每个视频以25帧/s的速率裁剪成图像帧，裁剪出来大约50张图片。为了便于处理，选择50帧为本次实验的基准帧，在数据读取的时候，如果图片的数量大于50张，程序会进行采样选取其中连续的50 帧图片；如果图片的数量小于50张，程序会随机选取部分图像帧进行复制，凑够50张图片。通过这样的处理，既保留了动态手势的核心运动信息，又避免了引入不必要的数据对实验结果产生影响。另外1280×720图片的尺寸太大，其中人做手语的部分太小，这样的图片输入网络中，一是会存在过多的无用信息，对神经网络拟合数据造成干扰；二是会增大计算开销，所以本文使用深度学习中目标检测的方法，切割出图像中有人的部分。但是每组图片中的人的大小是不一样的，这里运用Tensorflow中的image.resize_image(method)函数将图片大小统一调整为224×224。method参数选择双线性插值法。实验结果表明，经过处理之后的数据，对网络的精度大约有1.2%的提升。图5为RGB图像处理前后的对比。

3.3.2 光流图的提取

光流(optical flow)是视频动作识别分析的一种重要方法，它代表着3维物体的每个像素在像平面上运动的瞬时速度，一般来说光流也是物体在相邻两帧中运动变化情况的缩影。从光流信息中，不仅可以得到物体的运动方向、运动速度，还可以得到这个物体和我们的距离以及角度[21]。所以运用光流就可以很好地表达物体运动过程。得到光流图的方法有很多，本文使用的是基于Lucas-kanade方法的OpenCV中的calcOpticalFlowPyrLK函数。使用该函数成功提取每一帧的光流图并且保存，制作出单个单词的光流数据集。

3.4 实验过程

首先，使用连续50帧，每帧大小为224×224的RGB图像来训练网络。每一批次随机选取16个样本进行一次迭代。初始学习率为0.001，学习率衰

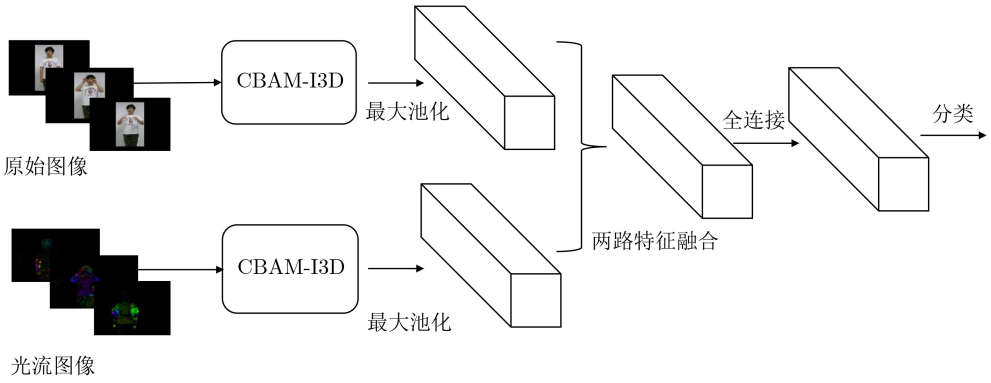
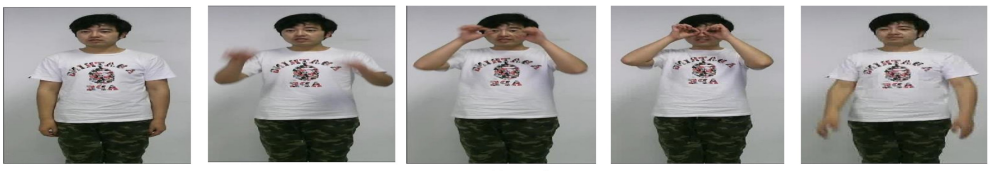


图 4 双流CBAM-I3D网络

(a) 处理前



(b) 处理后

图 5 RGB图像处理前后的对比

减为指数衰减，衰减系数为0.96，衰减步数为100。采用交叉熵损失函数，选用Adam Optimizer优化器对网络进行优化。共进行50000次迭代，在前1000次迭代中每100次保存1次训练权重，往后每隔1000次保存1次训练权重；然后用同样的方法训练一个光流图的模型，同样的方式保存权重。然后再将两个模型分别加载到双流网络中，以较低的学习率微调全部参数完成双流网络的训练。

3.5 实验结果分析

为了说明提出方法的有效性，本次实验对比了C3D网络、MFnet网络、3D ResNet网络和I3D网络在CSL数据集上的识别效果。对比结果如表1所示(其中平均检测时间是对于一段视频而言的)。C3D网络首先将3D-CNN应用于行为检测， 该网络的结构很简单，只有8个卷积层、5个池化层和2个全连接层，所以在CSL数据集上效果最差，耗时也最长。为了使3D卷积网络的计算量不至于过于庞大，MFnet被提了出来，MFnet通过分组卷积，降低了模型的计算量，MFnet的每一个分组都类似于ResNet的残差结构，但是一个输入会在通道维度上被分解为多个部分，所以从表1中能够看出虽然MFnet的Top1准确率不是很高，但是耗时却是最少的。为了更好地进行视频动作识别，研究者开始将深层网络如ResNet、Inception等2D网络扩展到3D网络，并对网络结构进行适当调整，从表1中也能看到，在CSL数据集中3D ResNet、I3D都取得了不错的效果，说明更深的网络结构在拟合复杂数据方面更有优势。另外从表1中也可以看出除了MFnet网络耗时时间较少外，其他网络耗时均较长，所以行为识别以后的研究方向在于网络达到高准确率的同时，也能实现低消耗。CBAM是一个轻量化的模块，可以完美地与其他网络进行融合，从表1中可以看出在添加了CBAM后，各网络都有不同程度的提升。另外双流网络对准确率也带来一定的提升。本文提出的CBAM-I3D网络在单流和双流的网络中都取得了的最好效果。实验结果验证了所提方法的有效性。

4 结束语

本文提出了一种融合双流3D卷积神经网络和注意力机制的新型动态手语识别网络，充分利用了RGB视频信息和光流信息对动态手势进行识别，在I3D网络中引入了CBAM，在不影响原始网络性能的同时，使网络学习到图像中的显著性信息，使得图像中重要的特征更加显著，提高了网络的表达能力，并且没有增加过多额外的参数和训练时间。同时使用BN层的标准化处理提高了网络的泛化性能，在加快训练效率的同时，还可以提升网络的性能。另外本文还对RGB视频信息进行了额外的处理，使用了目标检测的方法将人做手语的有用部分切割出来，增强了数据的可利用性，对网络的精度也有一定的提升。本文提出的方法不CBAM-I3D仅对手语识别有很好的效果，也完全可以应用于其他行为的识别。另外，本文提出的网络也存在一些不足，如网络层数比较深、3D卷积核参数比较多，

表 1 本文方法与其他方法在CSL数据集上的实验结果对比

网络结构	Top1准确率(%)	网络训练时间(h)	平均检测时间(s)
C3D(RGB)	70.53	35.15	2.05
CBAM-C3D(RGB)	71.77	36.26	2.10
C3D(RGB+Optical)	71.28	75.45	3.39
CBAM-C3D(RGB+Optical)	72.86	76.15	4.01
MFnet(RGB)	73.61	12.14	0.15
CABM-MFnet(RGB)	74.25	13.28	0.17
MFnet(RGB+Optical)	74.65	29.20	0.28
CABM-MFnet(RGB+Optical)	76.19	30.08	0.31
3D-ResNet(RGB)	79.52	24.42	0.58
CBAM-3D-ResNet(RGB)	82.90	25.43	0.61
3D-ResNet(RGB+Optical)	83.96	55.15	1.02
CBAM-3D-ResNet(RGB+Optical)	85.28	56.20	1.08
I3D(RGB)	84.56	20.29	0.41
CBAM-I3D(RGB)	86.00	21.31	0.42
I3D(RGB+Optical)	88.18	46.52	0.75
CBAM-I3D(RGB+Optical)	90.76	47.28	0.81

从而导致了网络的训练和优化比较困难，因此对网络结构进行优化将是未来的研究方向。

参 考 文 献

[1] TAKAHASHI T and KISHINO F. A hand gesture recognition method and its application[J]. *Systems and Computers in Japan*, 1992, 23(3): 38–48. doi: 10.1002/scj.4690230304.

[2] BANSAL B. Gesture recognition: A survey[J]. *International Journal of Computer Applications*, 2016, 139(2): 8–10. doi: 10.5120/ijca2016909103.

[3] 张淑军, 张群, 李辉. 基于深度学习的手语识别综述[J]. 电子与信息学报, 2020, 42(4): 1021–1032. doi: 10.11999/JEIT190416.

ZHANG Shujun, ZHANG Qun, and LI Hui. Review of sign language recognition based on deep learning[J]. *Journal of Electronics & Information Technology*, 2020, 42(4): 1021–1032. doi: 10.11999/JEIT190416.

[4] PARCHETA Z and MARTÍNEZ-HINAREJOS C D. Sign language gesture recognition using hmm[C]. The 8th Iberian Conference on Pattern Recognition and Image Analysis, Faro, Portugal, 2017: 419–426. doi: 10.1007/978-3-319-58838-4_46.

[5] PU Junfu, ZHOU Wengang, ZHANG Jihai, *et al*. Sign language recognition based on trajectory modeling with HMMs[C]. The 22nd International Conference on Multimedia Modeling, Miami, USA, 2016: 686–697. doi: 10.1007/978-3-319-27671-7_58.

[6] SAMANTA O, ROY A, PARUI S K, *et al*. An HMM framework based on spherical-linear features for online cursive handwriting recognition[J]. *Information Sciences*, 2018, 441: 133–151. doi: 10.1016/j.ins.2018.02.004.

[7] MASOOD S, SRIVASTAVA A, THUWAL H C, *et al*. Real-time sign language gesture (word) recognition from video sequences using CNN and RNN[M]. BHATEJA V, COELLO C A C, SATAPATHY S C, *et al*. Intelligent Engineering Informatics. Singapore: Springer, 2018: 623–632. doi: 10.1007/978-981-10-7566-7_63.

[8] DONAHUE J, JIA Yangqing, VINYALS O, *et al*. DeCAF: A deep convolutional activation feature for generic visual recognition[C]. The 31st International Conference on International Conference on Machine Learning, Beijing, China, 2014: I-647–I-655.

[9] TRAN D, BOURDEV L, FERGUS R, *et al*. Learning spatiotemporal features with 3d convolutional networks[C]. 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4489–4497. doi: 10.1109/ICCV.2015.510.

[10] CHEN Yunpeng, KALANTIDIS Y, LI Jianshu, *et al*. Multi-fiber networks for video recognition[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 364–380.

[11] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al*. Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: 10.1109/CVPR.2016.90.

[12] HUANG Jie, ZHOU Wengang, LI Houqiang, *et al*. Sign language recognition using 3D convolutional neural networks[C]. 2015 IEEE International Conference on Multimedia and Expo (ICME), Turin, Italy, 2015: 1–6. doi: 10.1109/ICME.2015.7177428.

[13] SIMONYAN K and ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. The 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014: 568–576.

[14] BAKER S, SCHARSTEIN D, LEWIS J P, *et al*. A database and evaluation methodology for optical flow[J]. *International Journal of Computer Vision*, 2011, 92(1): 1–31. doi: 10.1007/s11263-010-0390-2.

[15] CAO Zhe, SIMON T, WEI S E, *et al*. Realtime multi-person 2D pose estimation using part affinity fields[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 1302–1310. doi: 10.1109/CVPR.2017.143.

[16] CARREIRA J and ZISSERMAN A. Quo Vadis, action recognition? A new model and the kinetics dataset[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, USA, 2017: 4724–4733. doi: 10.1109/CVPR.2017.502.

[17] WOO S, PARK J, LEE J Y, *et al*. CBAM: Convolutional block attention module[C]. The 15th European Conference on Computer Vision, Munich, Germany, 2018: 3–19.

[18] HUANG Jie, ZHOU Wengang, ZHANG Qilin, *et al*. Video-based sign language recognition without temporal segmentation[C]. The 32nd AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, USA, 2018: 2257–2264.

[19] HU Jie, SHEN Li, and SUN Gang. Squeeze-and-excitation

networks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 2011–2023. doi: 10.1109/CVPR.2018.00745.

[20] IOFFE S and SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]. The 32nd International Conference on Machine Learning, Lille, France, 2015: 448–456.

[21] 刘天亮, 谯庆伟, 万俊伟, 等. 融合空间-时间双网络流和视觉注意的人体行为识别[J]. 电子与信息学报, 2018, 40(10): 2395–2401. doi: 10.11999/JEIT171116.

LIU Tianliang, QIAO Qingwei, WAN Junwei, *et al*. Human action recognition via spatio-temporal dual network flow and visual attention fusion[J]. *Journal of Electronics & Information Technology*, 2018, 40(10): 2395–2401. doi: 10.11999/JEIT171116.

王粉花：女，1971年生，博士，副教授，硕士生导师，研究方向为模式识别与智能信息处理.

张 强：男，1994年生，硕士生，研究方向为图像处理与手势识别.

黄 超：男，1993年生，硕士生，研究方向为图像处理.

责任编辑：马秀强