doi:10.3969/j.issn.1672-1292.2021.01.008

基于 3D-CBAM 注意力机制的人体动作识别

王　飞,胡荣林,金　鹰

(淮阴工学院计算机与软件工程学院,江苏 淮安 223003)

[摘要]　针对已有的动作识别方法的特征提取不足、识别率较低等问题,结合双流网络、3D 卷积神经网络和卷积 LSTM 网络的优势,提出一种融合模型. 该融合模型为了更好地提取人体动作特征,采用 SSD 目标检测方法将人体目标分割出作为局部特征和原视频的全局特征共同训练,并采用后期融合进行分类;将 3D 卷积块注意模块采用 shortcut 结构的方式融合到 3D 卷积神经网络中,加强神经网络对视频的通道和空间特征提取;并且通过将神经网络中部分 3D 卷积层替换为 ConvLSTM 层的方法,更好地得到视频的时序关系. 实验在公开的 KTH 数据集上进行. 结果表明,所提模型具有较高的人体动作识别准确率.

[关键词]　机器视觉,人体动作识别,3D 卷积神经网络,注意力机制
[中图分类号]TP391.4　[文献标志码]A　[文章编号]1672-1292(2021)01-0049-08

Human Action Recognition Based on 3D-CBAM Attention Mechanism

Wang Fei,Hu Ronglin,Jin Ying

(School of Computer and Software Engineering,Huaiyin Institute of Technology,Huaian 223003,China)

Abstract:Aiming at the problems of insufficient feature extraction and low recognition rate of existing action recognition methods,the paper proposes a fusion model by combining the advantages of two-stream network,3D convolutional neural network and convolutional LSTM network. In order to better extract human motion features,the fusion model adopts SSD target detection method to segment the human body as local features and global features of the original video for joint training,and adopts late fusion for classification. The 3D convolutional block attention module(3D-CBAM)is integrated into 3D convolutional neural network by using shortcut structure to enhance the neural network's channel and spatial feature extraction. And by replacing part of the 3D convolutional layer of the neural network with ConvLSTM layer,the temporal relation of the video is better obtained. The experiment is carried out on the KTH dataset,and the results show that the proposed model has high recognition accuracy of human action.

Key words:machine vision,human movement recognition,3D convolutional neural network,attention mechanism

人体动作识别是计算机视觉领域的一项基本任务,它基于一个视频中完整的动作执行来识别人体动作[1]. 随着机器学习研究的进行,人体动作识别的方法可以大致分为两种,一种是基于机器学习人工设计特征的方法[2-3],另一种是端到端的深度学习算法. 与人工制作的动作特征不同,深度学习方法可以从图像中自主学习特征,并且学习到的动作特征比人工动作特征有更好的识别性能. Tran 等[4]提出 C3D 网络模型,通过 3D 卷积直接从视频序列帧中提取时间和空间特征,但是仍然不能充分利用时间和空间特征. Simonyan 等[5]提出了双流卷积网络模型,这个模型由时间和空间网络构成,空间流从静态的视频序列帧中执行行为识别,同时时间流从密集光流形式的运动中训练以识别行为,但是双流网络需要人工提取出视频帧之间的光流信息以便时间流的训练识别. 文献[6-7]通过实验表明长短期记忆网络(long short-term memory,LSTM)能够一定程度上解决卷积神经网络(convolutional neural networks,CNN)不能够表示长时间变化的问题,但是 LSTM 是从 CNN 的全连接层获取特征进行处理的,所以缺乏时空特性的细节.

针对上述的问题,本文提出了一种在 C3D 卷积神经网络上进行改进的融合模型. 整个动作识别流程如图 1 所示. 首先通过 SSD 目标检测方法将视频中的人体目标进行分割,然后将原视频序列帧和分割后的人体目标序列帧经过一系列的 3D 卷积、ConvLSTM、3D 卷积块注意力模块(3D-convolutional block attention

收稿日期:2020-08-08.
通讯作者:胡荣林,博士,副教授,研究方向:人机交互技术. E-mail:huronglin@hyit.edu.cn

module,3D-CBAM)和 3D 池化模块直接提取时间和空间特征,最后将网络提取的特征进行后期融合并分类得到最终的人体动作分类结果.

1 相关研究

人体动作识别有着广泛的应用场景,如视频监控、视频的存储与检索、人机交互和身份识别等[8]. 人体动作识别涵盖了计算机视觉中的许多研究课题,包括视频中的人体检测、人体姿态估计、人体跟踪以及时间序列数据的分析和理解[9]. 随着深度学习在图像分类和目标检测上的成功应用,许多研究者也将其应用于人体动作识别. 与图像空间中的特征表示不同,视频中的人体动作表示不仅描述了人体在图像空间中的形象,而且还必须提取形象和姿态的变化即不仅需要提取外观信息还要提取运动信息. 目前,根据深度学习网络结构划分,可以将外观信息和运动信息结合的具有代表性的深度学习方法分为 3 种:基于双流卷积网络的方法、基于 3D 卷积网络的方法和基于长短期记忆网络的方法.

Simonyan 等[5]最先提出基于双流卷积网络的方法,首先对视频序列中每两帧计算密集光流,得到密集光流的序列,然后对视频图像和密集光流分别独立地训练 CNN 模型,两个网络分别对动作的类别进行判断,将两个分支网络的类别得分进行融合,最终得到动作分类结果. 文献[10-11]在双流网络的基础上利用 CNN 网络进行了空间和时间的融合,并将网络替换成 VGG-16,进一步提高了分类准确率. Wang 等[12]在使用 RGB 图像和光流图像的基础上还尝试了 RGB 差异和翘曲光流两种输入,通过实验证明 RGB、光流和翘曲光流的组合效果最好. 张聪聪等[13]将提取的关键帧融入双流卷积网络,相对降低网络复杂度并具有较高识别率. 基于双流网络的方法依赖提取的光流图像,然而光流的计算与存储代价比较昂贵.

Ji 等[14]最早提出 3D 卷积并将其运用到行为识别,提出的模型从空间和时间维度中提取特征,从而捕获在多个相邻视频帧中的运动信息,C3D 卷积网络是 3D 卷积网络的代表,通过 3D 卷积和 3D 池化可以对时间信息进行建模,并且可以将完整的视频帧作为输入,并不依赖于任何处理,可以轻松地扩展到大数据集. 但 3D 卷积仍然不能充分的提取时空特征.

LSTM 在时序数据上的处理能力比较强,但如果时序数据是图像,则在 LSTM 的基础上增加卷积操作,对于图像的特征提取会更加有效. Ng 等[15]采用 CNN 提取帧级特征,再将帧级特征和提取到的光流特征输入到 LSTM 进行训练得到分类结果. Shi 等[16]通过将全连接 LSTM 扩展为卷积结构,提出 ConvLSTM 网络,能够更好地捕捉时空相关性,并且始终优于 FC-LSTM 算法.

本文提出的融合模型综合利用了双流网络、3D 卷积网络和 ConvLSTM 网络 3 种网络的优势. 将双流网络的主体网络结构替换成 C3D,无需计算和存储光流信息,采用 ConvLSTM 层替代部分 3D 卷积层,提高对时间特征的利用率,能够更好地捕捉时空相关性. 同时在网络中采用 shortcut 结构[17]将 3D-CBAM 注意力机制结合到 3D 卷积中,提高了 C3D 卷积对空间特征的利用率.

2 研究方法

2.1 整体架构

本文的融合模型框架如图 1 所示. 首先对视频数据进行预处理,采用 SSD 目标检测方法对视频序列帧进行人体目标识别并分割. 然后将分割后的人体目标序列帧作为局部特征提取网络的输入数据用于提取局部运动特征,而原视频序列帧作为全局特征提取网络的输入数据用于提取全局运动特征. 最后将全局特征和局部特征融合并进行分类得到最终的人体动作分类结果.

2.2 局部特征提取

相机自身轻微的抖动和镜头的拉伸都会造成拍摄出的整个视频中存在运动信息,而这些运动信息并不是需要识别的人体运动信息,会影响神经网络训练的结果,所以局部信息的提取将会起到至关重要的作用. 本文采用 SSD 目标检测方法直接对视频序列帧的人体目标进行检测并分割作为局部信息. SSD 采用 VGG16 作为特征提取的主干网络,在 VGG16 的基础上新增了卷积层来获得更多的特征图以用于检测. SSD 采用多尺度特征图用于检测,比较大的特征图可以用来检测相对较小的目标,而小的特征图用来检测大目标. 借鉴文献[18]每个单元设置尺度或者长宽比不同的先验框,预测的边界框是以这些先验框为基准的,在一定程度上减少训练难度.

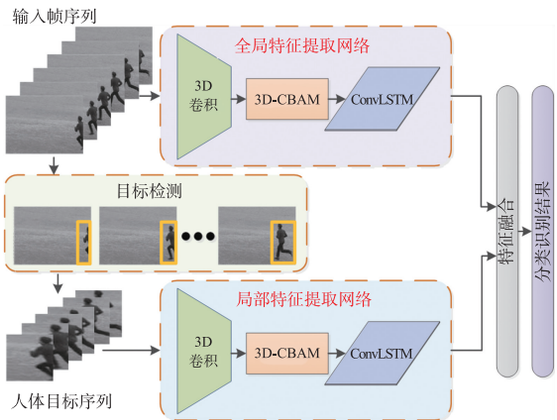


图 1　融合注意力机制的 3D 卷积网络动作识别框架

Fig. 1　3D convolutional network action recognition framework with attention mechanism

从卷积层中提取出作为检测所用的 6 个特征图,同一个特征图上每个单元设置的先验框是相同的,但不同的特征图设置的先验框数目是不同的. 先验框的设置需要确定其大小和长宽比. 先验框的大小随着特征图大小的降低成线性增加:

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1)\,k. \tag{1}$$

式中,$k \in [1,m]$,m 表示特征图个数;s_k 指先验框大小相对于图片的比例;$s_{\min}$ 和 $s_{\max}$ 表示先验框与图片比例的最小值和最大值. m 设置为 5,因为提取的第一个特征图是单独设置的,第一个特征图的先验框大小设置为 30. $s_{\min}$ 和 $s_{\max}$ 的值分别为 0.2 和 0.9. 对于长宽比,实验选取 1,2,3,1/2 和 1/3.

2.3　3D 卷积网络模型

本文的特征提取网络由两个部分组成:全局特征提取网络和局部特征提取网络. 这两个网络的基本框架相同,但对应输入的视频帧的维度不一样,全局特征提取网络的输入为 112×112×16×3,表示输入的视频帧大小为 112×112,每批次输入的视频帧数量为 16,通道为 3,而局部特征提取网络输入的视频帧大小为 64×64;全局特征提取网络中的 Pool 1 采用的最大池化窗口为 2×2×1,步长为 2×2×1,而局部特征提取网络中 Pool 1 采用的最大池化窗口为 2×2×2,步长为 2×2×2. 所有的 3D 卷积 Conv1,Conv2,…Conv5 都采用 3×3×3 大小的卷积核和 1×1×1 的步长,激活函数则使用 Relu 函数. 本文的融合模型在 C3D 网络的基础上进行改进,将原 C3D 网络中 Conv3b、Conv4b 和 Conv5b 层分别替换为 ConvLSTM1、ConvLSTM2 和 ConvLSTM3 层,并舍弃 FC7 层. 详细的网络结构模型如图 2 所示.

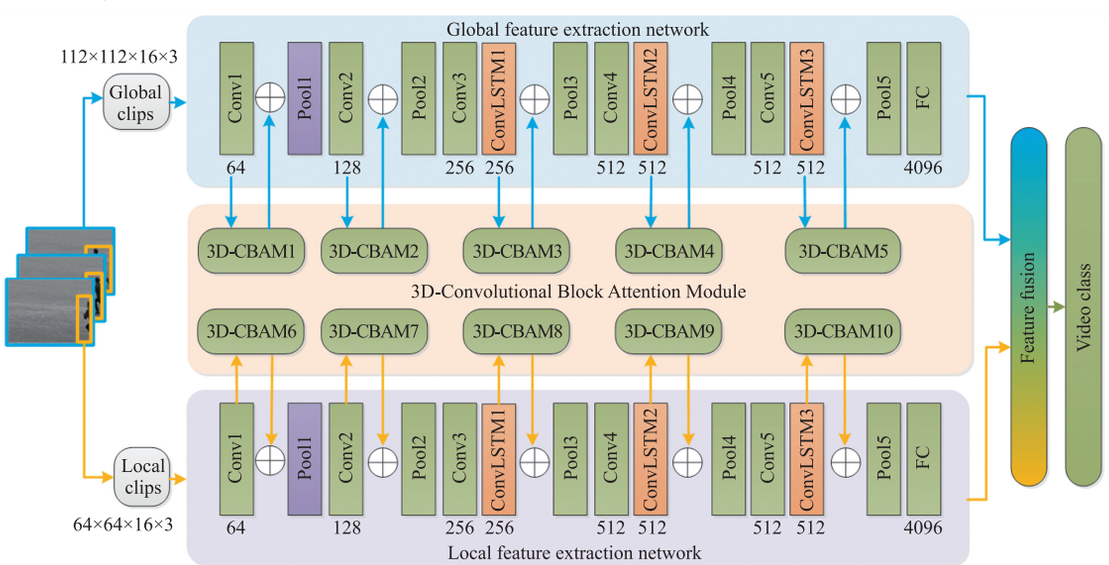


图 2　融合 3D-CBAM 注意力机制的 3D 卷积网络模型

Fig. 2　3D convolutional network model with 3D-CBAM attention mechanism fusion

与 LSTM 不同,ConvLSTM 的输入变换和循环变换是通过卷积实现的,即输入与各个门之间的连接、状态与状态之间由前馈式替换成卷积运算,ConvLSTM 的工作原理如下:

$$i_t=\sigma(W_{xi}*X_t+W_{hi}*H_{t-1}+W_{ci}\circ C_{t-1}+b_i),\tag{2}$$

$$f_t=\sigma(W_{xf}*X_t+W_{hf}*H_{t-1}+W_{cf}\circ C_{t-1}+b_f),\tag{3}$$

$$C_t=f_t\circ C_{t-1}+i_t\circ\tanh(W_{xc}*X_t+W_{hc}*H_{t-1}+b_c),\tag{4}$$

$$o_t=\sigma(W_{xo}*X_t+W_{ho}*H_{t-1}+W_{co}\circ C_{t-1}+b_o),\tag{5}$$

$$H_t=o_t\circ\tanh(c_t).\tag{6}$$

式中,$*$ 表示卷积操作,$\circ$ 表示乘积操作,$X_1,\cdots,X_t$ 为输入数据,$C_1,\cdots,C_t$ 为单元输出,$H_1,\cdots,H_t$ 为隐藏层,i_t、f_t 和 o_t 分别为网络中的输入门、遗忘门和输出门,W 和 b 分别表示对应门控的权重和偏置量,σ 为 sigmoid 激活操作.

本文融合模型使用的 ConvLSTM 层都选择 3×3 大小的卷积核,进行卷积操作时保留边界处的操作结果,对视频帧的所有像素点进行处理,使得输出的 shape 和输入的 shape 相同,采用 tanh 作为激活函数. ConvLSTM1 层将一个人体动作样本分为 8 个时间点输入即($X_1,X_2,\cdots,X_t$),此时 t 的值为 8,而 ConvLSTM2 层和 ConvLSTM3 层将一个人体动作样本分别分为 4 个和 2 个时间点输入,每个时间点都有相应的输出,所有的 ConvLSTM 层都是将所有时间点的结果输出并拼接作为整个 ConvLSTM 层的输出.

2.4　3D-CBAM 注意力机制

CBAM 注意力机制是可以直接应用于前馈卷积神经网络的简单而有效的注意模块,由通道注意模块和空间注意模块两个部分组成[19]. 对于卷积神经网络生成的特征图,CBAM 从通道和空间两个维度计算特征图的注意力图,然后将注意力图和特征图的对应元素相乘进行特征的自适应学习. CBAM 是一种轻量级的通用模块,目前研究者尝试过将其应用到诸如 VGG、Inception 和 ResNet 等 2D 卷积网络中进行端到端的训练. 本文为了提高 3D 卷积网络的空间特征利用率,提出 3D-CBAM 注意力机制,具体集成方式如图 3 所示.

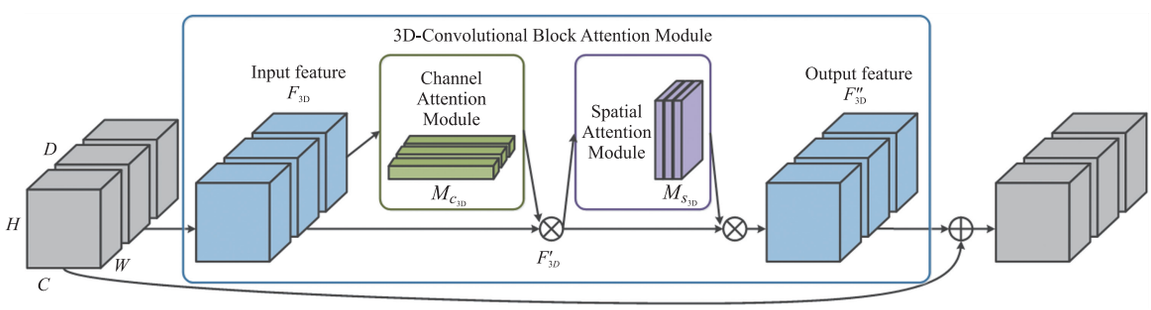


图 3　3D-CBAM 结构

Fig. 3　The structure of 3D-CBAM

与 2D 卷积网络不同的是 3D 卷积网络多出一个深度维度,在每一次提取空间特征和空间特征时,需要考虑到深度参数的变化. 对于一个中间 3D 卷积层的特征图:$F_{3D}\in\mathbf{R}^{W\times H\times D\times C}$,3D-CBAM 会按照顺序推理出通道注意力特征图 $M_{c_{3D}}\in\mathbf{R}^{1\times1\times1\times C}$,以及空间注意力特征图 $M_{s_{3D}}\in\mathbf{R}^{1\times H\times W\times D}$,整个过程公式如下所示:

$$F'_{3D}=M_{c_{3D}}(F_{3D})\otimes F_{3D}.\tag{7}$$

$$F''_{3D}=M_{s_{3D}}(F'_{3D})\otimes F'_{3D}.\tag{8}$$

3D-CBAM 的通道注意力模块关注哪些通道对融合 3D 网络的最后分类结果起到作用,即选择出对预测起决定性作用的特征,具体步骤如图 4 所示.首先将输入的特征图 F_{3D}分别经过基于宽度 W、高度 H 和深度 D 的最大值池化和均值池化,然后对分别经过 MLP 的特征进行基于对应元素的加和操作,再经过 sigmoid 激活操作,将生成的通道特征图 $M_{c_{3D}}(F_{3D})$与输入的特征图 F_{3D}进行相乘操作生成最终的通道特征图 F'_{3D},公式为:

$$M_{c_{3D}}(F_{3D})=\sigma(\mathrm{MLP}(\mathrm{AvgPool3D}(F_{3D}))+\mathrm{MLP}(\mathrm{MaxPool3D}(F_{3D}))=\sigma(W_1(W_0(F^c_{\mathrm{avg}}))+W_1(W_0(F^c_{\max}))).\tag{9}$$

式中,$W_0\in\mathbf{R}^{C/r\times C}$,$W_1\in\mathbf{R}^{C\times C/r}$,$\sigma$ 为 sigmoid 操作,W_0 需要经过 Relu 函数激活. 本文减少率 r 取值为 8,即在最大值池化和均值池化时将通道 C 变换为 $C/8$,减少参数量,最后再使用全连接变换为原来的通道 C.

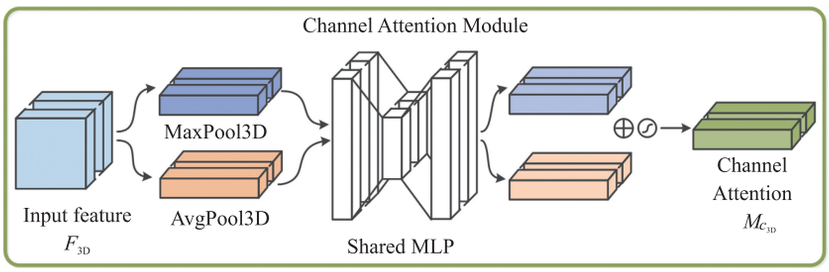


图 4　3D-CBAM 的通道注意力模块

Fig. 4　Channel attention module of 3D-CBAM

3D-CBAM 的空间注意力模型关注 RGB 图像中哪些位置的像素对网络的预测起到决定性作用,具体的注意力特征提取流程如图 5 所示. 首先将通道注意力模块的特征图 F'_{3D}作为空间注意力模块的输入特征图,对其做一个基于通道的最大值池化和均值池化操作,然后将提取出的两个特征 F^s_{avg}和 F^s_{max}进行基于通道的合并操作,接着通过一个 7×7 的卷积操作将其降维成一个通道再经过 sigmoid 激活函数生成空间注意力特征图,最后将该特征图与该模块输入的特征图 F'_{3D}进行对应元素的乘法操作得到最终生成的特征 F''_{3D},公式如下所示:

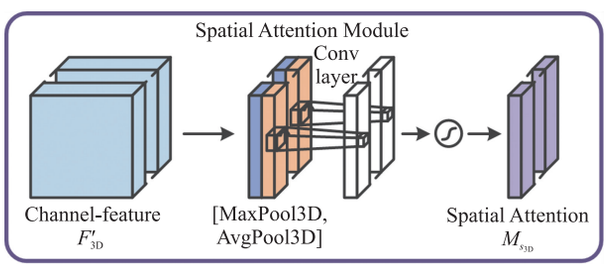


图 5　3D-CBAM 的空间注意力模块

Fig. 5　Spatial attention module of 3D-CBAM

$$M_{s_{3D}}(F'_{3D})=\sigma(f^{7\times7}([\mathrm{AvgPool3D}(F'_{3D});\mathrm{MaxPool3D}(F'_{3D})]))=\sigma(f^{7\times7}([F^s_{avg};F^s_{max}])). \tag{10}$$

实验证明,7×7 的卷积操作得到的实验效果优于 3×3 的卷积. 因为是应用于 3D 卷积且视频序列帧的通道排序格式为 channel-last,所以在进行合并操作时需要将张量中 axis=4 的通道串接,然后进行卷积操作保证 axis=4 的特征数为 1.

3　实验分析

3.1　数据集和评估指标

为了证明本文融合模型的有效性且考虑实验 GPU 运行内存等因素,在 KTH 数据集上进行实验,该数据集包含光照变换和相机自身运动的情况,贴合生活场景.

KTH 数据集为人类行动数据集,一共有 600 个视频,视频数据包含 6 种类型的人类动作,分别是步行、慢跑、奔跑、拳击、挥手和拍手. 由 25 个对象在 4 种不同的情况下进行的拍摄,分别是户外 s1、户外包含尺度变化 s2、户外穿着不同衣服 s3 以及室内 s4,如图 6 所示.

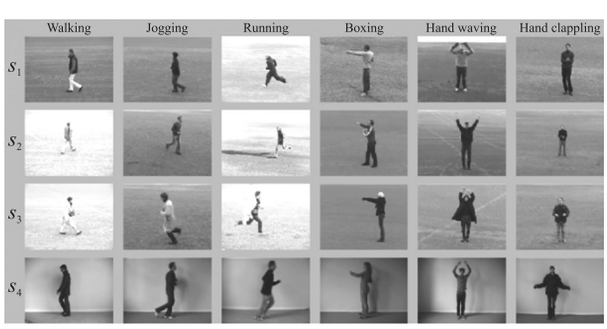


图 6　KTH 数据集示例图

Fig. 6　Sample diagram of KTH dataset

实验选择每个类别中 16 个对象的视频作为训练集,剩余的 9 个对象的视频作为验证集,每完成一次全部样本的训练就进行一次验证,总共进行 50 次操作,获得局部特征提取网络的权重和全局特征提取网络的权重,测试时调用模型权重提取特征并融合通过 SVM 分类获得最终的识别准确率. 本文将最终的识别准确率 A 作为动作识别的评估标准,公式如下:

$$A=\frac{N_{TP}+N_{TN}}{N_{TP}+N_{TN}+N_{FP}+N_{FN}}. \tag{11}$$

式中,N_{TP}为将正类样本预测为正类个数,N_{TN}表示将负类样本预测为负类个数,N_{FP}表示将负类样本预测为正类个数,N_{FN}为将正类样本预测为负类个数.

3.2　模型训练

3.2.1　数据处理

获取每一个视频的横纵比以及视频帧的总数,依次读取视频的每一帧,将每一帧由 BGR 格式转换为 RGB 格式,帧数据转换为数组形式并进行归一化处理. 加载 SSD300 的权值文件,利用其模型与权重对每一帧中的人体目标进行分割,最后将原视频帧和分割后的视频帧分别统一尺寸并保存,实验中原视频帧的大小为 112×112,分割后的视频帧大小为 64×64. SSD300 的权重文件是以 VOC 数据集训练的,能够检测 20 种目标,实验只需检测人体目标即可.

选取 16 人作为训练数据,每个视频选择 4 段连续帧,每段共有 16 帧,则训练数据集一共有 1 536 组连续帧,剩余的 9 人作为验证集,每个视频选择 1 段连续帧,同样是每段有 16 帧,一共有 215 组连续帧.

3.2.2　模型训练参数设置

在 Linux 系统搭建的 keras 平台下进行试验. 实验中使用训练集中的全部样本训练次数 epoch 为 50,考虑训练时 GPU 内存的情况,每一批次训练选取的样本数量为 10,实验训练优化器采用 SGD 随机梯度下降,SGD 使用 nesterov 动量,动量参数为 0.9,用于验证模型是否快速收敛;初始学习率 lr 为 0.005,训练时通过自定义回调函数的方法对学习率进行衰减,epoch 为 20、30 和 40 时学习率分别为 lr/10、lr/100 和 lr/1 000,在训练开始时使用较大的学习率可以使训练快速收敛,随着训练的过程逐渐降低学习率有助于找到最优解. 为了避免过拟合,对每一层卷积层使用 L2 正则化,并且在全连接层前采用值为 0.5 的 dropout.

3.2.3　模型测试

局部特征提取网络和全局特征提取网络训练后的结果都保存在二进制文件中,该文件包含模型的结构、模型的权重、训练配置(损失函数,优化器等)和优化器的状态. 通过迁移学习的方法将测试所用的样本经过文件存储的模型和权重处理获得测试集人体动作特征,然后将训练得到的局部特征和全局特征进行融合,采用(欧几里德)L2 范数对融合的特征进行归一化处理,最后用 SVM 进行特征分类获得最终的动作分类结果. 其中,SVM 是构建的软间隔分类器. 分类器的惩罚系数设置为 10,对边界内的噪声容忍度比较小,分类准确高;采用线性核函数进行计算分类;启用启发式收缩方式,能够预知哪些变量对应着支持向量,只需要在这些样本上进行训练即可,其他样本可以不予考虑,这种方式不仅不影响训练结果,还降低了问题的规模有助于迅速求解,起到一个加速的效果. 停止训练的误差精度设置为 0.001. 采用一对多法,即训练时依次把某个类别的样本归为一类,其他剩余的样本归为另一类,这样 n 个类别的样本就构造出 n 个分类器,分类时将未知样本分类为具有最大分类函数值的那类.

3.3　实验结果与分析

本文方法采用的是 3D 卷积网络对动作进行识别,相较于目前典型的双流网络减少了前期对视频的预处理操作,不需要单独提取出视频中的光流特征. 从表 1 中可以看出本文提出的方法与其他的方法相比取得更好的识别效果.

表 2 表明是否采用局部特征和全局特征融合的方法在动作识别中的准确率结果. 从表 2 中可以看出本文的方法无论是局部特征网络的准确率、全局特征网络的准确率还是两者融合后的动作识别准确率都比 C3D 网络和 ConvLSTM 网络的准确率高,同时局部特征与全局特征融合后的准确率比单独的全局特征的准确率高,这说明局部信息的提取弥补了单独全局特征提取的运动信息不足的缺点.

表 1　不同方法在 KTH 数据集上的动作识别准确率

Table 1　The action recognition accuracy of different methods on KTH dataset　%

方法	KTH
C3D	81.39
ConvLSTM	78.14
双流网络	80.93
FC-LSTM	78.07
本文方法	89.77

表 2　网络模型中局部特征和全局特征是否融合的动作识别准确率比较

Table 2　Accuracy comparison of action recognition based on fusion of local and global features or not in network model　%

方法	全局	局部	全局+局部
C3D	81.39	75.81	84.65
ConvLSTM	78.14	80.47	84.18
本文方法	84.19	83.72	89.77

为了证明 3D-CBAM 注意力机制在本文的融合模型中的有效性,实验分别对其是否使用注意力机制进行了对比实验,表 3 展示了是否使用 3D-CBAM 注意力机制的动作识别准确率的结果. 由表 3 可以看出

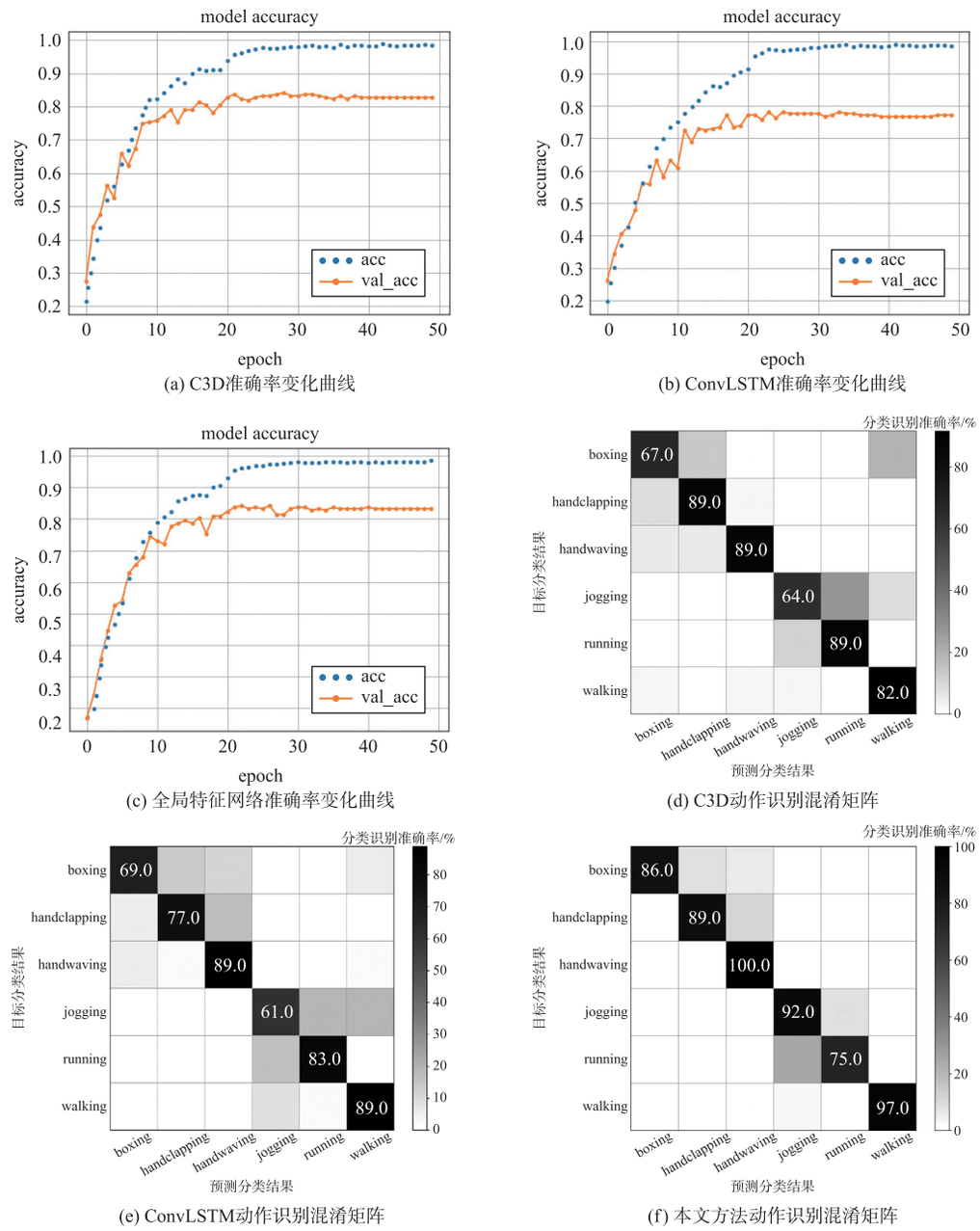
3D-CBAM 注意力机制的添加使得 C3D 网络和 ConvLSTM 网络对动作识别的准确率都得到了显著的提升.本文方法使用 3D-CBAM 注意力机制比不使用注意力机制的动作识别准确率高,该实验证明了 3D-CBAM 注意力机制对人体动作识别任务的有效性.

图 7 为部分实验训练过程和测试结果的可视化展示. 因为本文方法中的局部特征网络和全局特征网络是单独训练的,且后期融合采用的是 SVM,所以仅采用全局特征网络准确率变化曲线与其他方法对比,可以体现出主体网络的优势. 图 7(a)、图 7(b)和图 7(c)分别为 C3D、ConvLSTM 和本文融合模型中全局特征网络准确率变化曲线,点线为训练数据的准确率变化曲线,折线则是验证数据集的准确率变化曲线,很明显可以看出本文的融合模型的全局特征提取网络在验证集上的准确率要高于 ConvLSTM 网络,虽然与 C3D 网络的准确率相近,但是在迭代 10 次

表 3　3D-CBAM 注意机制使用与否的动作识别准确率比较

Table 3　Comparison of motion recognition accuracy of 3D-CBAM in use or not　%

方法	不使用 3D-CBAM	使用 3D-CBAM
C3D	81.39	83.25
ConvLSTM	78.14	79.07
本文方法	87.91	89.77



(a) C3D准确率变化曲线　(b) ConvLSTM准确率变化曲线

(c) 全局特征网络准确率变化曲线　(d) C3D动作识别混淆矩阵

(e) ConvLSTM动作识别混淆矩阵　(f) 本文方法动作识别混淆矩阵

图 7　实验结果数据可视化

Fig. 7　Data visualization of experimental results

内的波动范围更小,说明可以更好的学习时空特征,3 种网络都是迭代次数在 20~30 次之间达到最高精确度. 图 7(d)、图 7(e)和图 7(f)为测试时 C3D、ConvLSTM 和本文融合模型的混淆矩阵热图. 从混淆矩阵热图中可以看出,标签为 handwaving 的准确率本文提出的融合方法已经达到 100%,除了标签为 running 的准确率相比于其他两种方法低,其余的标签的准确率都有明显的提高,最少提高了 8%,最多提高了 31%.

4　结论

针对人体动作识别现有方法的优缺点,本文提出了一种融合模型. 该模型在 C3D 网络的基础上加入了 ConvLSTM 模块并融合了 3D-CBAM 注意力机制,通过局部特征提取网络和全局特征提取网络提取出局部特征和全局特征并进行融合提高动作识别的准确率. 实验在 KTH 数据集上进行,实验结果表明该模型对于人体动作识别能够达到很好的识别效果. 本文虽然采用轻量级的 3D-CBAM 注意力机制,但如何缩减参数量和计算量仍然是需要进一步研究和解决的问题.

[参考文献](References)

[1] KONG Y,FU Y. Human action recognition and prediction:a survey[EB/OL]. [2020-05-21]. https://arxiv.org/abs/1806.11230.

[2] WANG H,KLASER A,SCHMID C,et al. Action recognition by dense trajectories[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs,USA,2011.

[3] WANG H,SCHMID C. Action recognition with improved trajectories[C]//Proceedings of the IEEE international conference on computer vision. Sydney,Australia,2013.

[4] TRAN D,BOURDEV L,FERGUS R,et al. Learning spatio-temporal features with 3D convolutional networks[EB/OL]. [2020-05-20]. https://arxiv.org/abs/1412.0767.

[5] SIMONYAN K,ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems. Montréal,Canada,2014.

[6] SRIVASTAVA N,MANSIMOV E,SALAKHUTDINOV R. Unsupervised learning of video representations using LSTMs[EB/OL]. [2020-06-11]. https://arxiv.org/abs/1502.04681v3.

[7] FRANCISCO O,DANIEL R. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition[J]. Sensors,2016,16(1):115-140.

[8] 罗会兰,童康,孔繁胜. 基于深度学习的视频中人体动作识别进展综述[J]. 电子学报,2019,47(5):1162-1173.

[9] ZHANG H B,ZHANG Y X,ZHONG B,et al. A comprehensive survey of vision-based human action recognition methods[J]. Sensors,2019,19(5):105-120.

[10] FEICHTENHOFER C,PINZ A,ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition,Las Vegas,USA,2016.

[11] SIMONYAN K,ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2015-05-20]. https://arxiv.org/abs/1409.1556.

[12] WANG L,XIONG Y,WANG Z,et al. Temporal segment networks:towards good practices for deep action recognition[C]//European Conference on Computer Vision. Cham:Springer,2016.

[13] 张聪聪,何宁. 基于关键帧的双流卷积网络的人体动作识别方法[J]. 南京信息工程大学学报(自然科学版),2019,64(6):96-101.

[14] JI S W,XU W,YANG M,et al. 3D convolutional neuralnetworks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):221-231.

[15] NG Y H,HAUSKNECHT M,VIJAYANARASIMHAN S,et al. Beyond short snippets:deep networks for video classification[EB/OL]. [2020-05-21]. https://arxiv.org/abs/1503.08909.

[16] SHI X,CHEN Z,WANG H,et al. Convolutional LSTM network:a machine learning approach for precipitation nowcasting[C]//Advances in Neural Information Processing Systems. Montreal,Quebec,Canada,2015.

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

[18] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37(2):142-158.

[19] WOO S,PARK J,LEE J Y,et al. CBAM:convolutional block attention module[C]//European Conference on Computer Vision. Munich,Germany,2018.

[责任编辑:陈　庆]