DOI: 10.13382/j.jemi.B2104402

基于分割注意力的特征融合 CNN-Bi-LSTM 人体行为识别算法*

余金锁 卢先领
(江南大学物联网工程学院 无锡 214122)

摘 要:针对传统人体行为识别算法不能有效抑制空间背景信息,网络间缺乏信息交互,以及无法对全局时间相关性进行建模的问题,提出一种基于分割注意力的特征融合卷积神经网络-双向长短时记忆网络(CNN-Bi-LSTM)人体行为识别算法。首先以一定采样率采样 30 帧图像,通过分割注意力网络提取图像的深度特征,并引入特征融合机制增强不同卷积层间的信息交互;然后将深度特征输入到 Bi-LSTM 网络对人体动作的长时时间信息建模,最后使用 Softmax 分类器对识别结果进行分类。相较于传统双流卷积网络,该算法在 UCF101 和 HMDB51 数据集上的准确率分别提高了 6.6%和 10.2%,有效提高了识别准确率。

关键词: 行为识别;分割注意力;特征融合;双向长短时记忆网络
中图分类号: TP391 **文献标识码**: A **国家标准学科分类代码**: 520.6040

Human action recognition algorithm of feature fusion CNN-Bi-LSTM based on split-attention

Yu Jinsuo Lu Xianling
(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China)

Abstract: Aiming at the problems that traditional human action recognition algorithms cannot effectively suppress spatial background information, the lack of information interaction between networks, and the inability to model global temporal correlation, a human action recognition algorithm of feature fusion Bi-LSTM based on segmentation attention is proposed. First, 30 frames of images are sampled at a certain sampling rate, extract the depth features of the images by split-attention network, and introduce a feature fusion mechanism to enhance the information interaction between different convolutional layers. Then input the depth features into the Bi-LSTM network to model the long-term information of human actions, and finally use the Softmax classifier to classify the recognition results. Compared with the traditional two-stream convolutional network, the accuracy of this algorithm on the UCF101 and HMDB51 datasets is increased by 6.6% and 10.2%, respectively, which effectively improves the recognition accuracy.

Keywords: action recognition; split-attention; feature fusion; BI-LSTM

0 引 言

人体行为识别是指利用模式识别、机器学习等方法,自动分析识别出视频中的人体行为,在视频检索,安全监控,智能家居等领域有着广泛应用[1-2]。与单图像识别相比,动作识别不仅需要提取动作的空间特征,还要对视频帧之间的时间相关信息进行建模。此外视频中的背景、光照、视角等变化也会影响识别效果,因此设计一个高效准确的行为识别算法十分具有挑战性。

传统人体行为识别算法通过提取运动目标的特征描述子[3-5]等人工特征来表征人体动作,从而达到识别的目的。其中最具代表性的算法是密集轨迹算法(dense tnajectory, DT)[6]和改进的密集轨迹算法(improved dense trajectory, iDT)[7]。然而传统方法所提取的人工特征受限于人的经验和动作的单一性,鲁棒性和迁移能力不强。

近年来,深度学习发展迅速,卷积神经网络(convolutional neural networks, CNN)在图像领域取得了巨大成功。使用深度学习中的卷积神经网络通过自学习

收稿日期:2021-06-08 Received Date: 2021-06-08
*基金项目:国家自然科学基金(61773181)项目资助

的方式来提取动作特征,成为人体行为识别领域的主流方法。文献[8]提出使用独立的空间流和时间流分别提取视频中的空域信息和时域信息,构成双流卷积神经网络。双流网络极大地提高了识别准确率,但是双流网络在空间流仅操作一帧,在时间流仅操作短片段中的单堆帧,对视频中的时间信息利用有限。文献[9]据此提出时序分割网络(temporal segment networks, TSN),通过将整段视频稀疏地采样为多个视频片段,分别对这些片段进行预测,最后对每个片段得分进行融合。时序分割网络充分利用视频的时空信息,有效实现了对长视频的建模。文献[10]针对二维卷积无法提取时间信息的缺点,将卷积神经网络中的卷积层和池化层扩展到三维,提出C3D 网络。C3D 网络使用大小为 3×3×3 的卷积核直接在时间和空间维度操作 16 帧视频输入,最后使用 Softmax 分类器得到视频分类结果。C3D 网络仅使用 RGB 视频帧作为输入,不需要计算额外的光流,提升了数据预处理速度,但是由于其三维结构,网络参数也成倍增加。文献[11]将 C3D 中的 3×3×3 卷积核分解为 1×3×3 和 3×1×1,分别用来提取空间和时间信息。该方法可以有效减少 3D 卷积的计算量,提高模型训练速度。

循环神经网络(recurrent neural network, RNN)由于其出色的时间建模能力,也被广泛应用于行为识别领域。通过使用 CNN 提取视频帧的动作特征,再利用 RNN 的变体之一长短期记忆网络(long short-term memory, LSTM)[12]对动作的时间信息建模,可以有效提高识别精度,并减少工作量。但是 CNN 提取的动作特征会直接影响最后的识别结果。为了更好地提取视频帧中的运动信息,并对视频帧之间的长时时间信息建模,本文提出了基于分割注意力的特征融合 CNN-双向 LSTM(CNN-Bi-LSTM)网络模型。该模型利用卷积神经网络提取动作特征,并引入分割注意力机制[13],使其在学习动作的抽象信息时,更加关注通道间的信息交互[14-16]。其次,通过融合不同卷积层特征,实现信息的互补,提高特征表达能力[17-18]。卷积神经网络提取的动作特征通过全局平均池化和全连接层输出尺寸为 1×1×2 048 的一维向量,并输入到 Bi-LSTM 中以建模时间信息,最后使用 Softmax 分类器得到每个动作的最后得分。

1 算法框架

1.1 分割注意力网络

为解决空间背景等冗余信息对网络所提取特征精度的影响,提出使用分割注意力网络(ResNeSt)提取动作的空间特征。分割注意力网络继承了残差网络(residual network, ResNet)[19]简单且模块化的优点,并引入通道注意力机制,在不增加网络复杂度的同时提升特征图精确性,用该网络提取动作特征可以兼顾准确率和效率的平衡。

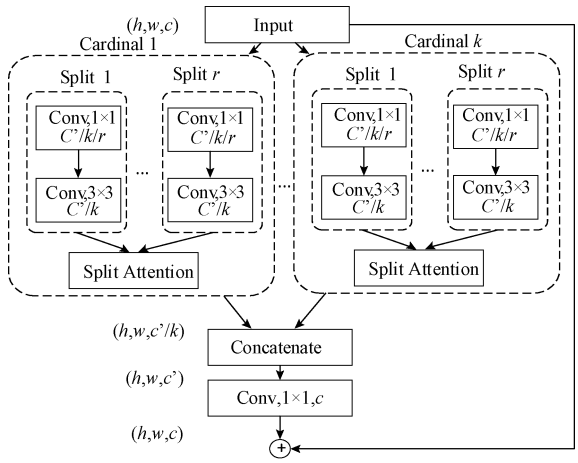

图 1 ResNeSt 模块结构

Fig. 1 ResNeSt block structure

如图 1 所示,在网络模块中,首先将输入分成 k 个基数组,再将每个基数组分成 r 个分割,那么特征图组的总数可以记为 $G = kr$。对每一个特征图组使用通道维度的注意力,以提高通道间的交互性,通过这种细化分组并赋予通道权重的方式来提升特征细粒度。每个分组的中间表示为:

$$U_i = F_i(X), i \in \{1,2,\cdots,G\} \tag{1}$$

式中:F_i 是对每个独立分组使用的 1×1 卷积和 3×3 卷积。全局上下文信息 S^k 是嵌入在通道中的全局统计信息,可以描述视频帧中像素间的关联。在分割注意力模块中通过使用全局上下文信息 S^k 计算不同分组在通道维度的权重 $\boldsymbol{a}_i$。第 c 通道的全局上下文信息 S_c^k 计算公式如下:

$$S_c^k = \frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{U}_c^k(i,j) \tag{2}$$

式中:H、W 为特征图的输出大小;$\hat{U}_c^k(i,j)$ 表示第 c 个通道的特征向量卷积后的结果。利用全局上下文信息 S_c^k 计算得到的特征图组权重 $\boldsymbol{a}_i$ 可表示为:

$$a_i^k(c) = \begin{cases} \dfrac{exp(G_i^c(s^k))}{\sum_{j=0}^{r} exp(G_j^c(s^k))}, r > 1 \\ \dfrac{1}{1 + exp(-G_j^c(s^k))}, r = 1 \end{cases} \tag{3}$$

式中:权重函数 G_i^c 是根据全局上下文信息 S^k 确定的每个分组在第 c 个通道所占的权重,该权重函数使用 BN 层进行归一化操作,并使用带有 ReLu 函数激活的全连接层初始化。

每个通道的特征图由不同特征图组加权组合产生,基于通道维度软注意力实现的第 c 个通道特征图可表

示为：

$$V_c^k = \sum_{i=1}^{r} a_i^k(c)\, U_{r(k-1)+i} \tag{4}$$

最后，对每一个特征图组沿通道维度进行连接，得到分割注意力模块的输出：

$$V = Concat\{V^1, V^2, \cdots, V^K\} \tag{5}$$

将原始输入 x 与分割注意力块的输出 V 进行 shortcut 短接，从而生成残差注意力模块输出：

$$Y = V + T(x) \tag{6}$$

式中：T 表示带步长的卷积操作，用于对齐分割注意力块的输出形状。

1.2　多层特征融合

卷积神经网络中，低层特征分辨率高，具有丰富的位置，轮廓信息，但其噪声更多，语义信息更少。随着网络深度提升，图像分辨率不断降低，特征图中高级语义信息更丰富却缺乏空间信息，对细节感知能力较差。多层特征融合考虑卷积神经网络不同层次特征间的区别，通过平均池化和拼接等操作融合二者，提高信息的交互，从而提升人体行为识别算法准确率。本文以 ResNeSt 为骨干网络，并在网络中引入特征融合机制以提高 CNN 特征的表征能力，以下为融合细节。

ResNeSt 网络主要分为 5 个模块，每个模块的输出如表 1 所示。

表 1　ResNeSt 网络各层输出尺寸

Table 1　Output size of each layer of ResNeSt

层	输出尺寸
Conv_custom	128×56×56
Layer1	256×56×56
Layer2	512×28×28
Layer3	1 024×14×14
Layer4	2 048×7×7

在 Conv_custom 层中使用 3 个 3×3 卷积核替代 7×7 卷积核进行卷积，在不改变感受野的情况下减少了计算量。在 Conv_custom 层最后使用最大池化，输出尺寸为 128×56×56 的特征图。再经过一系列卷积和池化等操作，在 Layer4 层输出 2 048×7×7 的特征图输出。ResNeSt 网络在前向卷积的同时，将 Layer1 层的输出通过平均池化操作，使尺寸降低到 7×7，作为浅层特征与 Layer4 层融合。这里选择 Layer1 层输出作为浅层特征是因为 Layer1 层相较于 Conv_custom 层，该层经过了更多卷积操作，去除了冗余的背景和噪声，相对于 Layer2、Layer3 层又保留了更多的空间位置信息。经过池化后的浅层特征与 Layer4 层的深层特征在通道维度进行叠加（concatenation），以达到融合特征的目的。之后再次使用全局平均池化将融合后的特征展开至 1×1×2 048，并将该一维向量传输至 Bi-LSTM 网络中。

1.3　Bi-LSTM

传统 RNN 网络在反向传播时，容易发生梯度消失和梯度爆炸问题，难以从长时输入序列中学习时序特征。RNN 网络变体之一，LSTM 网络通过引入“门”结构来更新“细胞状态”，有效缓解了梯度消失问题。其单元结构如图 2 所示。

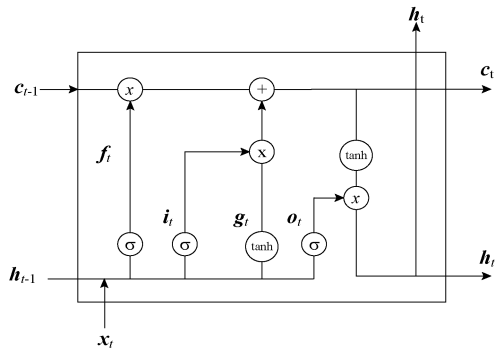

图 2　LSTM 单元结构

Fig. 2　LSTM unit structure

$$\boldsymbol{i}_t = \sigma(\boldsymbol{U}_i\boldsymbol{x}_t + \boldsymbol{W}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{7}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{U}_f\boldsymbol{x}_t + \boldsymbol{W}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{8}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{U}_o\boldsymbol{x}_t + \boldsymbol{W}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{9}$$

$$\boldsymbol{g}_t = \tanh(\boldsymbol{U}_g\boldsymbol{x}_t + \boldsymbol{W}_g\boldsymbol{h}_{t-1} + \boldsymbol{b}_g) \tag{10}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t\boldsymbol{g}_t \tag{11}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \times \tanh(\boldsymbol{c}_t) \tag{12}$$

式(7)～(12)为 LSTM 单元处理时间序列数据的过程。其中 $\boldsymbol{h}_{t-1}$、$\boldsymbol{h}_t$ 分别表示前一时刻和当前时刻的输出；$\boldsymbol{c}_{t-1}$ 和 $\boldsymbol{c}_t$ 用于纪录细胞前一时刻和当前时刻状态；$\boldsymbol{i}_t$、$\boldsymbol{o}_t$ 是输入门和输出门，用于控制信息的流入和流出，遗忘门 $\boldsymbol{f}_t$ 用于筛选先前序列中有用信息。$\boldsymbol{u}_i$、$\boldsymbol{u}_f$、$\boldsymbol{u}_o$、$\boldsymbol{u}_g$ 以及 $\boldsymbol{w}_i$、$\boldsymbol{w}_f$、$\boldsymbol{w}_o$、$\boldsymbol{w}_g$ 为前一时刻特征向量和当前时刻特征向量经过控制门的权重，通过反向传播迭代更新；$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_g$ 为偏置项。

LSTM 网络非常适合处理序列数据，然而其单向结构并未利用序列数据的未来信息。Bi-LSTM 网络对其进行修改，在前向传递的同时增加反向传递结构。该结构在处理视频数据时不但能够利用之前的视频帧，还依赖于之后的视频帧，充分利用人体动作的前后相关特性。Bi-LSTM 结构如图 3 所示。

$$\boldsymbol{h}_t = \sigma(\boldsymbol{w}_1\boldsymbol{x}_t + \boldsymbol{w}_2\boldsymbol{h}_{t-1} + \boldsymbol{b}_t^{(1)}) \tag{13}$$

$$\boldsymbol{h}_t' = \sigma(\boldsymbol{w}_3\boldsymbol{x}_t + \boldsymbol{w}_5\boldsymbol{h}_{t+1}' + \boldsymbol{b}_t^{(2)}) \tag{14}$$

$$\boldsymbol{O}_t' = \tanh(\boldsymbol{w}_4\boldsymbol{h}_t + \boldsymbol{b}_t^{(3)}) \tag{15}$$

$$\boldsymbol{O}_t'' = \tanh(\boldsymbol{w}_6\boldsymbol{h}_t' + \boldsymbol{b}_t^{(4)}) \tag{16}$$

$$\boldsymbol{O}_t = \frac{(\boldsymbol{O}_t' + \boldsymbol{O}_t'')}{2} \tag{17}$$

式(13)～(17)为 BI-LSTM 网络处理时间序列数据

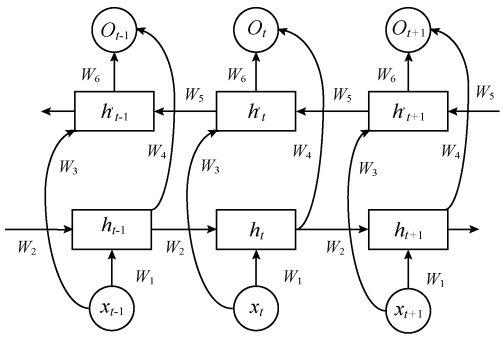

图 3 Bi-LSTM 网络结构

Fig. 3 Bi-LSTM networks structure

的过程。式中:$\boldsymbol{w}_i$(i=1, …,6)表示权重,$\boldsymbol{b}_t$ 为偏置项;$\boldsymbol{x}_i$(…, $\boldsymbol{x}_{t-1}$, $\boldsymbol{x}_t$, $\boldsymbol{x}_{t+1}$, …)为卷积神经网络提取的特征,以时间顺序依次输入 Bi-LSTM 中;$\boldsymbol{o}_t$ 为时序输出向量,取前向传播与反向传播的平均值作为输出。与传统单向 LSTM 相比,Bi-LSTM 利用了序列数据的双向时间信息,提高了信息利用率。

2 实验过程及结果分析

2.1 数据集及预处理

UCF101 数据集是佛罗里达大学发布的最具挑战性的人体行为识别数据集之一。该数据集收集于 YouTube 视频网站,主要包括体育运动、乐器演奏和人物交互等 101 类动作,共有 13 320 个视频,每个视频平均时长约为 7 s。HMDB51 数据集是由布朗大学发布的人体行为识别数据集,该数据集主要涵盖面部动作,身体动作,人物交互等 51 类 6 849 个视频数据。由于该数据集视频数量少、质量较差、背景复杂,使该数据集更具挑战性。在这两个数据集提供的 3 个训练测试分割上进行实验,并取 3 个分割的平均值作为最终实验结果。其中 70%用于训练,30%用于测试。此外本文还在 2017 年发布的大型数据集 Something-something 上进行了实验,该数据集包括 174 种动作类别的 108 499 个视频片段,由 80%的训练视频和 20%的测试视频组成。

分别在上述 3 个数据集上进行训练和测试,以验证本文所提出算法的有效性。将数据集压缩包解压后,逐帧分解每个视频得到视频帧,并保存到对应文件夹中。视频帧的原始尺寸为 320×240,调整为 224×224 后输入到网络以进行实验。

2.2 实验环境

本文实验在戴尔 PowerEdge T640 服务器上完成,运行环境为 Intel(R) Xeon(R) Silver 4110 CPU,GPU 为 Tesla P100,显存 16 GB,服务器操作系统为 Ubuntu 16.04,使用 Pytorch 深度学习框架,CUDA 版本为 10.0。训练过程中使用小批量随机梯度下降算法学习网络参数,动量设置为 0.9,批训练大小为 64,初始学习率设置为 1×10^{-3},每 20 个训练周期将学习率调整为原来的 10%,共训练 100 个周期。

2.3 自适应采样策略

主流人体行为识别数据集中存在视频长短不一现象,导致不同视频用于行为识别的视频帧数量不一定。对视频帧进行针对性采样至关重要,不同长度视频采用不同采样率采样,从而最大程度上利用视频帧中的时间信息。本文提出自适应视频长度的视频帧采样策略,对于视频帧小于 60 帧的输入,对其前 30 帧进行密集采样;视频帧介于 60~90 帧的输入,以采样率 2 对其采样 30 帧;对大于 90 帧的输入,以采样率 3 对 9 其采样 30 帧。需要注意的是,采样率不宜过大,过大会导致运动信息丢失,降低识别准确性。

2.4 实验结果分析

实验结果以识别准确率进行评价,即识别正确的数量与视频总数的百分比。实验过程中,以 ResNet101-BI-LSTM 网络作为基线,采用自适应采样策略在 UCF101 数据集和 HMDB51 数据集上进行消融实验,分别验证分割注意力网络和特征融合模块对识别结果的影响。实验结果如表 2 所示。

表 2 在 UCF101、HMDB51 数据集上的消融实验结果

Table 2 Results of ablation experiments on UCF101 and HMDB51 data sets

消融模型	参数量/M	准确率/%	
		UCF101	HMDB51
ResNet101-BI-LSTM	45.4	89.6	63.8
ResNeSt101-BI-LSTM	48.9	93.4	68.3
Ours	48.9	94.6	69.6

如表 2 所示,使用原始的 ResNet101-BI-LSTM 网络在 UCF101 和 HMDB51 数据集上的准确率分别为 89.6%和 63.8%。在引入分割注意力机制后,模型参数量增加了 3.5 M,识别准确率分别提高了 3.8%和 4.5%,识别效果明显提升。说明 CNN 在提取动作特征过程中,分割注意力机制在增加少量模型参数的情况下,能够有效地区分动作发生时的前景与背景,提高通道间信息的交互,提升识别精确度。

随着卷积层数的不断增加,浅层特征所携带的空间、轮廓信息逐渐消失,取而代之的是更为抽象的高级语义信息。将 Layer1 层与 Layer4 层的特征进行融合之后,识别准确率分别达到了 94.6%和 69.6%,相较于融合前提升了 1.2%和 1.3%。说明融合浅层特征所携带的空间信息与深层特征的高级语义信息,可以提高特征图的表征能力,从而有效提高识别准确率。此外,由于特征融合过

程中使用的是池化等操作,所以模型参数量并未明显提升。

为了更直观地表示本算法在 UCF101 和 HMDB51 数据集上的表现,制作其识别结果的混淆矩阵图,如图 4 所示。其中,混淆矩阵的横轴表示预测的动作类别,纵轴表示实际动作类别。对角线位置表示识别正确率,对角线以外的方块表示误识别为其他动作的概率。方块颜色表示概率大小,具体如右侧柱状图所示。

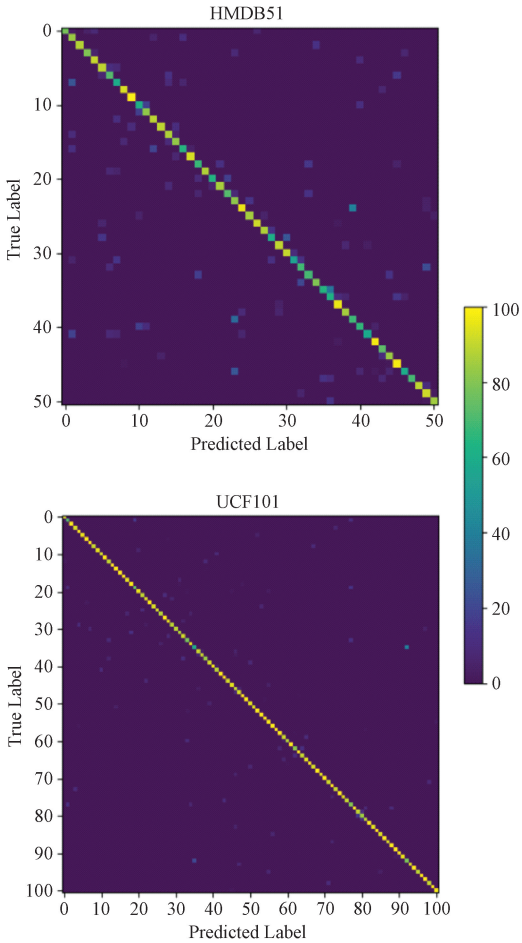

图 4　HMDB51 和 UCF101 数据集准确率混淆矩阵

Fig. 4　Accuracy confusion matrix on HMDB51 and UCF101dataset

根据混淆矩阵可以看出,对于背景及运动轨迹相似的动作,如 UCF101 数据集中的涂口红和剃胡子,扔链球和扔飞饼,以及 HMDB51 数据集中的笑和微笑等高度相似动作误识别率较高。对于其他绝大部分动作能够准确识别其动作类型,误识别率低。

表 3 为本文所提出方法与现有主流方法的识别结果对比。与传统手工制作特征方法相比,本文所提出的算法准确率在 UCF101 和 HMDB51 数据集上分别提高了 6.7%和 8.5%;相较于经典深度学习方法 TwoStream,本文所提出方法准确率提高了 6.6%和 10.2%。出于平衡模型参数量与识别精度的考虑,本文使用的骨干网络为 ResNeSt101,这导致了本文方法参数量略大于部分对比方法。因此,为进一步验证所提出方法的有效性,本文在大型数据集 Something 上进行了对比实验,结果如表 3 所示。与 ECO 方法相比,本文方法参数量高出 1.4 M,但是本文方法在 3 个数据集上的识别准确率要高出其 1%左右;相较于 ECO_{En},本文方法的参数量仅为其 1/6,但是也取得了相似的精度。综合来看,本文方法在模型参数量与识别精度上达到了较好的平衡。

表 3　不同算法实验结果对比

Table 3　Comparison of experimental results of different algorithms

方法	参数量/M	准确率/%		
		UCF101	HMDB51	Someth.
IDT+HSV[6]	–	87.9	61.1	–
Res3D[20]	33.3	85.8	54.9	–
TwoStream[8]	12	88.0	59.4	–
TDD[21]	117.6	90.3	63.2	–
3D Conv[22]	79	91.8	64.6	–
R(2+1)D[23]	63.6	93.6	66.6	–
ECO[24]	47.5	92.8	68.5	41.4
ECOEn[24]	300	94.8	72.4	43.9
Ours	48.9	94.6	69.6	42.8

3　结　论

人体行为识别是一项具有挑战性的任务,具有广泛的应用前景。针对传统算法在提取动作特征时,通道间信息交互性不强以及特征利用率低的问题,本文提出了一种基于分割注意力的 CNN-Bi-LSTM 人体行为识别算法。该算法使用分割注意力网络提取视频中动作的表观信息,有效抑制了视频帧中冗余的背景,将目光集中于动作本身。同时使用特征融合机制,提升高低层特征间的交互。最后利用 Bi-LSTM 网络对动作特征进行时域建模,并通过 Softmax 分类器得到视频各类别概率,达到动作识别目的。实验结果表明,该算法在 UCF101、HMDB51 以及 Something 数据集上的准确率分别为 94.6%、69.6%和 42.8%,有效提高了算法准确率,验证了所提方法的有效性。本文仅利用 RGB 图像作为输入,避免了复杂的光流图提取和训练过程,这也导致了部分相似动作识别不理想的情况。在后续工作中会重点关注光流图,时序动态图等多模态输入下的识别情况,以及模型复杂度与识别准确率间的平衡。

参考文献

[1] 罗会兰, 王婵娟, 卢飞. 视频行为识别综述[J]. 通信学报, 2018, 39(6): 169-180.
LUO H L, WANG CH J, LU F. Survey of video behavior

recognition [J]. Journal on Communications, 2018, 39(6): 169-180.

[2] 王丽君, 刘彦戎, 王丽静. 基于卷积长短时深度神经网络行为识别方法[J]. 电子测量与仪器学报, 2020, 34(9): 160-166.
WANG L J, LIU Y R, WANG L J. CLDNN based human activity recognition method [J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(9): 160-166.

[3] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005: 886-893.

[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, 2008: 1-8.

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

[6] WANG H, SCHMID C. Dense trajectories and motion boundary descriptors for action recognition [J]. International Journal of Computer Vision, 2013, 103(1):60-79.

[7] WANG H, SCHMID C. Action recognition with improved trajectories[C]. Proceedings of the IEEE International Conference on Computer Vision, 2013: 3551-3558.

[8] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]. Advances in Neural Information Processing Systems, 2014: 548-475.

[9] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition [C]. European Conference on Computer Vsion, 2016: 20-36.

[10] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[11] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks [C]. Proceedings of IEEE International Conference on Computer Vision, 2017: 5534-5542.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] ZHANG H, WU C R, ZHANG Z Y, et al. ResNeSt: Split-attention networks [C]. Computer Vision and Pattern Recognition, 2020, arXiv: 2004.08955.

[14] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[15] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 510-519.

[16] XIE S, GIRSHICK R, PIOTR D, et al. Aggregated residual transformations for deep neural networks [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1492-1500.

[17] YU W, YANG K, YAO H, et al. Exploiting the complementary strengths of multi-layer CNN features for image retrieval [J]. Neurocomputing, 2016, 237: 235-241.

[18] 李洪均, 丁宇鹏, 李超波, 等. 基于特征融合时序分割网络的行为识别研究[J]. 计算机研究与发展, 2020, 57(1): 145-158.
LI H J, DING Y P, LI CH B, et al. Action recognition of temporal segment network based on feature fusion[J]. Journal of Computer Research and Development, 2020, 57(1): 145-158.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[20] DU T, RAY J, SHOU Z, et al. ConvNet architecture search for spatiotemporal feature learning[J]. Computer Vision and Pattern Recognition, 2017, arXiv:1708.05038.

[21] WANG L, QIAO Y, TANG X. Action recognition with trajectory-pooled deep-convolutional descriptors [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4305-4314.

[22] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013: 221-231.

[23] DU T, HENG W, LORENZO T, et al. A closer look at spatiotemporal convolutions for action recognition [C]. IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6450-6459.

[24] ZOLFAGHARI M, SINGH K, BROX T, et al. ECO: Efficient convolutional network for online video understanding [C]. IEEE Conference on European Conference on Computer Vision, 2018:713-730.

作者简介

余金锁,2018 年于安徽工业大学获得学士学位,现为江南大学硕士研究生,主要研究方向为图像处理与模式识别。
E-mail:yjs26597@163.com

Yu Jinsuo received B. Sc. degree from Anhui University of Technology in 2018. Now he is a M. Sc. candidate at Jiangnan University. His main research interests include image processing and pattern recognition.

卢先领(通信作者),2009 年于南京理工大学获得博士学位,现为江南大学物联网工程学院教授,主要从事无线传感器网络、移动边缘计算方向研究。
E-mail:jnluxl@jiangnan.edu.cn

Lu Xianling (Corresponding author) received his Ph. D. degree from Nanjing University of Science and Technology in 2009. He is currently a professor at Jiangnan University. His main research interests include wireless sensor networks and mobile edge computing.