基于 Bi-LSTM-Attention 模型的人体行为识别算法

朱铭康，卢先领[2]*

[1]江南大学轻工过程先进控制教育部重点实验室，江苏 无锡 214122；
[2]江南大学物联网工程学院，江苏 无锡 214122

摘要 针对长短时记忆网络(LSTM)不能有效地提取动作前后之间相互关联的信息导致行为识别率偏低的问题，提出了一种基于 Bi-LSTM-Attention 模型的人体行为识别算法。该算法首先从每个视频中提取 20 帧图像，通过 Inceptionv3 模型提取图像中的深层特征，然后构建向前和向后的 Bi-LSTM 神经网络学习特征向量中的时序信息，接着利用注意力机制自适应地感知对识别结果有较大影响的网络权重，使模型能够根据行为的前后关系实现更精确的识别，最后通过一层全连接层连接 Softmax 分类器并对视频进行分类。通过 Action Youtobe 和 KTH 人体行为数据集与现有的方法进行比较，实验结果表明，本文算法有效地提高了行为识别率。

关键词 机器视觉；行为识别；注意力机制；Inceptionv3 模型；长短时记忆网络

中图分类号 TP391 **文献标识码** A **doi**：10.3788/LOP56.151503

Human Action Recognition Algorithm Based on Bi-LSTM-Attention Model

Zhu Mingkang[1], Lu Xianling[2]*

[1] *Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi, Jiangsu* 214122, *China*;
[2] *School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu* 214122, *China*

Abstract This study proposed a human action recognition algorithm based on the Bi-LSTM-Attention model to solve the problem of low action recognition rate. This problem was caused by the inability of long short term memory (LSTM) networks to effectively extract correlative informations before and after actions. The proposed algorithm first extracted 20 image frames from each video and used the Inceptionv3 model to extract deep features from these frames. Then, forward and backward Bi-LSTM neural networks were constructed to learn the temporal information in the feature vectors. The influences of network weights on recognition results were adaptively perceived using the attention mechanism. This step was performed so that the model could achieve more accurate recognition based on the relationship between informations acquired before and after performing the given action. Finally, the videos were connected via a fully-connected layer to a Softmax classifier for classification. Comparison between the Action Youtobe and KTH human action datasets and existing methods revealed that the proposed algorithm effectively improved the action recognition rate.

Key words machine vision; action recognition; attention mechanism; Inceptionv3 model; long short term memory networks

OCIS codes 150.1135; 100.3008; 100.4996; 100.5010

1 引 言

随着视频监控、智能家居、无人驾驶的广泛应用，行为识别逐渐成为计算机视觉领域的研究热点。早期的人体行为识别大多使用人工设计的方法提取视频特征[1-2]。文献[3-4]提出了一种基于时空兴趣点的行为识别方法，该方法通过计算每个位置的强弱信息来标记视频中突出的部分。Abdulmunem 等[5]提出三维尺度不变特征变换-定向光流(3D-SIFT-HOOF)特征提取的人体行为识别方法，该方

收稿日期：2019-01-23；**修回日期**：2019-02-19；**录用日期**：2019-03-11

基金项目：教育部-新华三集团“云数融合”基金(2017A13055)

* **E-mail**：jnluxl@jiangnan.edu.cn

法通过提取 3D SIFT 和 HOOF 特征，并使用可视化语言方法对其进行编码。Luo 等[6]使用稀疏编码的方法学习静态特征，并用基于最大池化的时域金字塔结构对特征进行直方图表示，最后采用支持向量机(SVM)进行分类。Liu 等[7]提出了一种分层聚类多任务学习(HC-MTL)方法，通过目标函数加强特定行为的特征来实现行为识别。基于手工特征提取的方法在行为识别方面取得了许多优异的成果[8]，然而，人工设计的方法由于动作的多样性通常容易忽略一些重要的特征，对于行为识别的结果有着较大的影响。

近年来随着深度学习的快速发展，基于深度神经网络的特征提取方法逐渐取代了人工提取特征的过程[9-10]。Ji 等[11]首次提出了一种 3D 卷积神经网络(3D CNN)算法，该方法通过对时间轴上的视频帧运用 3D 卷积核来捕捉时间轴上视频的空间与时间信息，并将其用来识别人体行为。Donahue 等[12]提出了一种长期循环卷积神经网络(LRCN)，该网络从 CNN 中提取特征并通过长短时记忆网络(LSTM)来学习这些特征的顺序关系。Gammulle 等[13]提出了一种利用卷积层输出来训练 LSTM 的人体行为识别算法，将最终卷积层输出与第一个全连接层输出的两种 LSTM 流相结合进行视频分类。李庆辉等[14]通过 Rank 支持向量机算法将光流序列压缩成有序光流图，最后将双流网络的 C3D 描述子和 VGG 描述子融合后输入到线性 SVM 进行行为识别。Das 等[15]利用深度时域的 LSTM 结合外观和运动的深度特征，利用相同的输入特征对时域进行建模，并利用特征选择机制对 CNN 进行分类。Ullah 等[16]利用 CNN 提取视频帧的深层特征，并利用双向 LSTM 学习特征序列中的时序信息，最后通过 softmax 分类器进行分类。在行为识别中，CNN 和 LSTM 的使用极大地提高了识别精度，并减少了工作量。但是 CNN 的深度对视频帧的特征提取有着较大的影响：低层次的网络模型不易表现出图像的深度特征，容易出现欠拟合；深层次的网络模型容易产生梯度消散，难以优化网络模型。LSTM 无法有效地学习运动的时序特征，缺乏自主适应能力。

综上，为了更好地提取运动视频的特征信息，并增强 LSTM 时序特征学习能力。本文设计了一种 Bi-LSTM-Attention 模型，该模型在使用 Inceptionv3 模型[17]增加卷积神经网络深度的同时减少网络参数，充分提取视频帧的深度特征，然后把每帧相应的特征向量传入 Bi-LSTM 网络中以充分学习视频帧之间的时序特征，再把特征向量传入到 Attention 层中自适应地感知对识别结果有较大影响的网络权重，使一些特征能够得到更多的关注，最后通过全连接层连接 Softmax 分类器得到最终的分类结果，用于识别人体行为。

2 基于 Bi-LSTM-Attention 模型的行为识别

CNN 模型适用于二维的图像处理，通过对图片进行卷积池化操作提取其深度特征。如果把视频的每一帧单独看作一张图片进行处理，将极大降低网络模型的学习效率。因此本文针对每个视频取 20 帧作为样本。然后把样本传入到 Inceptionv3 模型中提取深度特征。接着把特征向量送到本文模型中进行网络权重学习，以达到最优的识别效果。本文的总体框架如图 1 所示，共分为三个部分，分别为 Inceptionv3 层、Bi-LSTM 层、Attention 层。其中 Bi-LSTM 层、Attention 层组成 Bi-LSTM-Attention 模型，为本文算法的核心部分。

1) Inceptionv3 层

本文的 Inceptionv3 层主要是对输入的视频帧进行特征提取，即把这些视频帧处理成 Bi-LSTM 层能够直接接收并能处理的特征向量形式。与传统的 CNN 特征提取方法不同，Inceptionv3 模型通过不同的卷积核对图像进行卷积操作，然后将不同的卷积层通过并联的方式结合在一起，最后再把得到的特征向量融合起来，部分结构如图 2 所示。图中的 $128\times128\times3$ 代表本文视频帧大小(128×128 代表像素，3 代表通道数)，1×1、$1\times n$、$n\times1$ 代表卷积核大小，pool 代表池化层操作。最后通过 filter concatenate 把不同卷积核处理的特征向量拼接起来，通过全连接层输出 $S\times1024$(S 代表视频帧个数，本文为 20)维的深度特征矩阵并将其传输到 Bi-LSTM 层。

2) Bi-LSTM 层

Bi-LSTM 是由一个向前传播和一个向后传播的 LSTM 组成。在 LSTM 结构中通过记忆控制器来决定遗忘和保留哪些信息，通过输入门、遗忘门和输出门三个结构来实现信息的输入和输出，单元结构如图 3 所示，图中：$\boldsymbol{x}_t$(t 代表视频帧，数值在 0～20 之间)代表输入的特征向量；$\boldsymbol{f}_t$ 表示通过遗忘门后被舍弃的信息；$\boldsymbol{c}_{t-1}$ 表示更新前的单元，$\boldsymbol{c}_t$ 表示更新后的单元，其中 $\boldsymbol{c}$ 表示记忆控制向量，它决定上一特征向量需要保存到下一特征向量的信息；

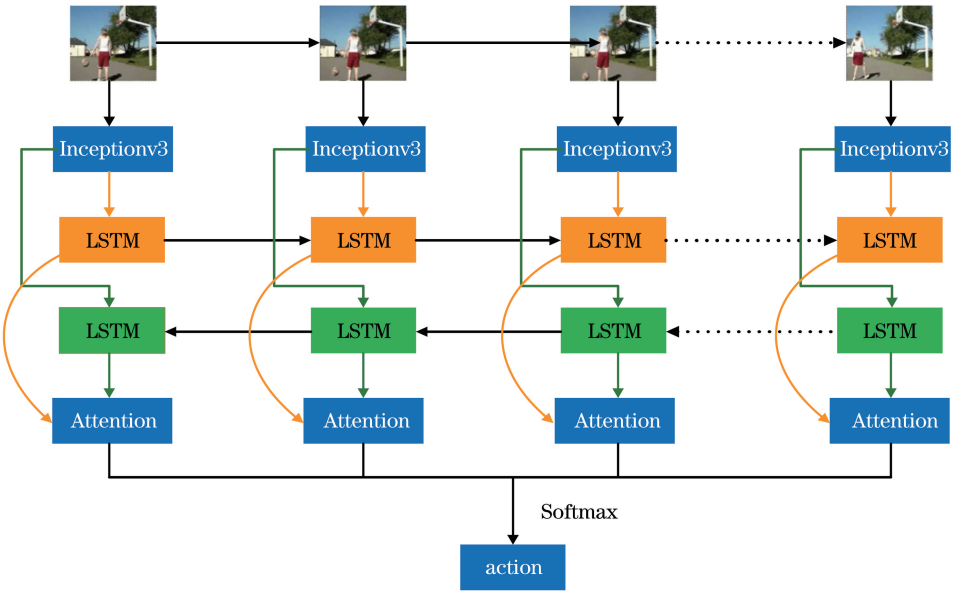

图 1 基于 Bi-LSTM-Attention 模型的行为识别框架

Fig. 1 Action recognition framework based on Bi-LSTM-Attention model

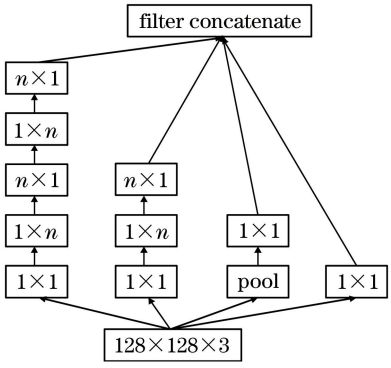

图 2 Inceptionv3 部分结构示意图

Fig. 2 Partial structural diagram of Inceptionv3

$\boldsymbol{h}_{t-1}$和 $\boldsymbol{h}_t$ 分别表示上一特征向量的输出和当前特征向量的输出；$\boldsymbol{i}_t$ 表示输入门状态更新率，它表示需要更新的信息；$\boldsymbol{g}_t$ 表示状态更新向量；$\boldsymbol{p}_t$ 表示当前输入值对应的输出；sigmoid 和 tanh 代表激活函数。运算过程表示为

$$\boldsymbol{f}_t=\text{sigmoid}(\boldsymbol{w}_{\text{f}}\boldsymbol{h}_{t-1}+\boldsymbol{u}_{\text{f}}\boldsymbol{x}_t+\boldsymbol{b}_{\text{f}}),\tag{1}$$

$$\boldsymbol{i}_t=\text{sigmoid}(\boldsymbol{w}_{\text{i}}\boldsymbol{h}_{t-1}+\boldsymbol{u}_{\text{i}}\boldsymbol{x}_t+\boldsymbol{b}_{\text{i}}),\tag{2}$$

$$\boldsymbol{g}_t=\tanh(\boldsymbol{w}_{\text{g}}\boldsymbol{h}_{t-1}+\boldsymbol{u}_{\text{g}}\boldsymbol{x}_t+\boldsymbol{b}_{\text{g}}),\tag{3}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\boldsymbol{g}_t,\tag{4}$$

$$\boldsymbol{p}_t=\text{sigmoid}(\boldsymbol{w}_{\text{o}}\boldsymbol{h}_{t-1}+\boldsymbol{u}_{\text{o}}\boldsymbol{x}_t+\boldsymbol{b}_{\text{o}}),\tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{p}_t\times\tanh(\boldsymbol{c}_t),\tag{6}$$

式中：$\boldsymbol{w}_{\text{f}}$、$\boldsymbol{w}_{\text{i}}$、$\boldsymbol{w}_{\text{g}}$、$\boldsymbol{w}_{\text{o}}$ 以及 $\boldsymbol{u}_{\text{f}}$、$\boldsymbol{u}_{\text{i}}$、$\boldsymbol{u}_{\text{g}}$、$\boldsymbol{u}_{\text{o}}$ 为上一特征向量的输出和当前特征向量的输入经过每一个控制门的权重；$\boldsymbol{b}_{\text{f}}$、$\boldsymbol{b}_{\text{i}}$、$\boldsymbol{b}_{\text{g}}$、$\boldsymbol{b}_{\text{o}}$ 为通过控制门的偏置项。(1)式计算通过遗忘门后被舍弃的信息。(2)、(3)式通过输出门计算状态更新率 $\boldsymbol{i}_t$ 和状态更新向量 $\boldsymbol{g}_t$ 的值，然后给定输入门状态更新率 $\boldsymbol{i}_t$，遗忘门激活值 $\boldsymbol{f}_t$ 和状态更新向量 $\boldsymbol{g}_t$，通过(4)式计算 $\boldsymbol{c}_{t-1}$ 在此 LSTM 单元的更新值 $\boldsymbol{c}_t$。(5)、(6)式通过输出门决定单元状态的哪一部分将被输出。

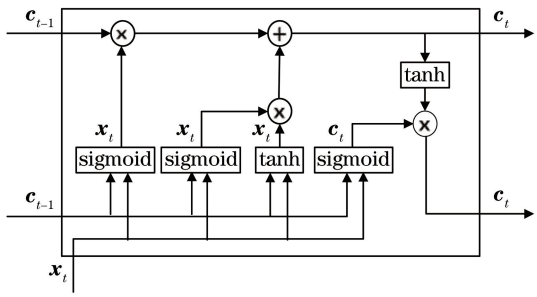

图 3 LSTM 单元结构

Fig. 3 LSTM cell structure

传统的 LSTM 网络只能单向学习，从而忽略了未来的信息。在 Bi-LSTM 中，当前时刻的输入不仅依赖之前的视频帧，还依赖于之后的视频帧，两个单元相互结合充分考虑了视频帧前后的时序信息，模型结构如图 4 所示。图中：$\boldsymbol{w}_i$($i=1,\cdots,6$)表示一个单元层到另一单元层的权重，$\boldsymbol{x}_t$ 表示所取得视频帧通过 Inceptionv3 层提取深层特征后得到的特征向量(1×1024)，$\boldsymbol{h}$ 表示$\{\cdots,\boldsymbol{x}_{t-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1},\cdots\}$从前向后输入的特征序列组成的 LSTM 单元，$\boldsymbol{h}'$表示$\{\cdots,\boldsymbol{x}_{t+1},\boldsymbol{x}_t,\boldsymbol{x}_{t-1},\cdots\}$从后向前输入的特征序列组成的 LSTM 单元，$\boldsymbol{o}_t$ 表示特征向量通过 Bi-LSTM 网络后对应的输出结果。

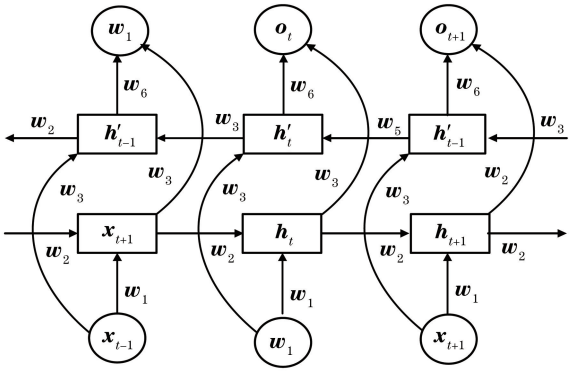

图 4 Bi-LSTM 网络模型

Fig. 4 Bi-LSTM network model

$$\boldsymbol{h}_t=\mathrm{sigmoid}(\boldsymbol{w}_1\boldsymbol{x}_t+\boldsymbol{w}_2\boldsymbol{h}_{t-1}+\boldsymbol{b}_t^{(1)}),\tag{7}$$

$$\boldsymbol{h}'_t=\mathrm{sigmoid}(\boldsymbol{w}_3\boldsymbol{x}_t+\boldsymbol{w}_5\boldsymbol{h}'_{t+1}+\boldsymbol{b}_t^{(2)}),\tag{8}$$

$$\boldsymbol{o}'_t=\tanh(\boldsymbol{w}_4\boldsymbol{h}_t+\boldsymbol{b}_t^{(3)}),\tag{9}$$

$$\boldsymbol{o}''_t=\tanh(\boldsymbol{w}_6\boldsymbol{h}'_t+\boldsymbol{b}_t^{(4)}),\tag{10}$$

$$\boldsymbol{o}_t=\frac{(\boldsymbol{o}'_t+\boldsymbol{o}''_t)}{2},\tag{11}$$

式中：$\boldsymbol{b}_t^{(1)}$、$\boldsymbol{b}_t^{(2)}$、$\boldsymbol{b}_t^{(3)}$、$\boldsymbol{b}_t^{(4)}$ 为第 t 个特征向量在 Bi-LSTM 网络中隐藏单元的偏置；$\boldsymbol{o}'_t$、$\boldsymbol{o}''_t$ 为两个 LSTM 单元在相应时刻分别处理 Inceptionv3 层输出的特征向量的结果。如(11)式所示，将相应时刻的两个特征向量相加求和取平均值的结果作为输出特征向量 $\boldsymbol{o}_t$，最后再把特征向量送入到注意力机制中进行感知网络权重。与传统的单项 LSTM 算法相比，Bi-LSTM 算法可以同时学习过去和将来的信息，从而获得更加稳健的时间信息。

3) Attention 层

Attention 机制是一种类似人类视觉所特有的大脑信号处理机制，通过计算不同时刻从 Bi-LSTM 网络中输出的特征向量的权重，突出一些重要特征，使整个网络模型能够表现出更好的性能。在行为识别中，神经网络在训练模型时通过添加 Attention 机制来重点关注一些关键动作和物体，例如打篮球，其中的跳、抬手动作和篮球、篮筐物体会被 Attention 机制分配更多的权重用来加深模型记忆，当模型下次遇到这几类动作或物体时会优先对这些行为做出预测，缩小识别的范围，然后再根据动作之间的关系调整权重的分配以达到更准确的识别。本文的注意力模型如图 5 所示。其中 $\boldsymbol{o}_t$ 表示从 Bi-LSTM 网络中输出的第 t 个特征向量，把特征向量传送到注意力机制模型中，经过注意力模型中的隐藏层得到初始状态向量 $\boldsymbol{s}_t$(1024×1)。权重系数 α_t 表示初始输入状态向量在最终输出的状态向量 $\boldsymbol{Y}$ 中所占的比重。对各初始输入的状态向量 $\boldsymbol{s}_t$ 与权重系数 α_t 的乘积进行累加求和得到最终输出的状态向量 $\boldsymbol{Y}$。计算公式为

$$e_t=\tanh(\boldsymbol{w}_t\boldsymbol{s}_t+b_t),\tag{12}$$

$$\alpha_t=\frac{\exp(e_t)}{\sum_{j=0}^{t}e_j},\tag{13}$$

$$\boldsymbol{Y}=\sum_{t=0}^{19}\alpha_t\boldsymbol{s}_t,\tag{14}$$

式中：b_t 为一个能量偏置；e_t 为第 t 个特征向量的状态向量 $\boldsymbol{s}_t$ 所决定的能量值。根据(13)式，利用以 e 为底数各部分能量值的次方与之前部分的能量值的累加和的比值可以得到对识别结果有影响的权重系数，由此实现了初始状态到注意力状态的转换，然后通过(14)式得到最终输出的状态向量 $\boldsymbol{Y}$，最后将 $\boldsymbol{Y}$ 通过全连接层整合在一起作为一个输出值，减少特征位置对分类带来的影响，利用 Softmax 分类器将多个神经元的输出映射到(0,1)区间内，从而进行多分类。

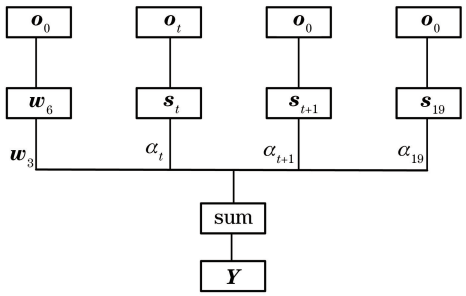

图 5 注意力机制模型

Fig. 5 Attention mechanism model

3 实验与分析

3.1 实验数据集

ActionYoutobe 数据集采集于 Youtobe 网站，包含 11 个动作类别，分别是投篮、骑自行车、跳水、高尔夫挥杆、骑马、踢足球、荡秋千、网球挥杆、蹦床、排球扣球以及遛狗。每种类别包含 25 组视频，每组有 4 个以上的视频片段，同一组视频具有一些共同的特征，例如相同的背景、同一执行者和相同位置的视角。所有类别一共有 1600 个视频，每个视频帧的大小都是 320 pixel×240 pixel。

KTH 数据集由固定摄像机视角拍摄的 600 个动作视频组成，包含了尺度变化、衣着变化和光照变化，视频帧的大小为 160 pixel×120 pixel。该数据集包括在 4 个不同场景下 25 人完成的 6 类动作：走、慢跑、快跑、挥手、拍手、拳击，实验者在每个环境下每个动作各执行 1 次。

3.2 实验设置

本实验使用 python 语言在 GPU 加速环境下进行实验，采用 keras 深度学习框架，电脑配置为 Win10 系统、16 GB 内存、GTX1080 11G 显存。由于 Inceptionv3 模型训练的数据类别与数量与本实验有很大差别，因此通过对 Inceptionv3 模型进行微调，屏蔽其最后 5 层，然后去训练 Bi-LSTM-Attention 模型中的网络参数。表 1 列举了本实验的参数设置。

表 1 实验参数

Table 1 Experimental parameters

Parameter	Value
Loss function	Categorical_crossentropy
Optimizer	Adam
Learning rate	0.0001
Batch_size	16
Epoch	100

为了验证 Bi-LSTM-Attention 模型的有效性，选取 Action Youtobe 数据集中的 60%作为训练集来训练网络模型，20%作为验证集来评估模型的性能，剩下的 20%作为测试集。每次取 KTH 数据集中的 80%进行训练，剩下的 20%进行测试，由于数据集较少，进行 5 次交叉验证取平均值并与当前的方法进行对比，具体划分情况如表 2 所示。

表 2 数据集划分

Table 2 Dataset division

Dataset	Training	Validation	Test	Cross validation
Action Youtobe	960	320	320	0
KTH	480	0	120	5

3.3 实验评价指标

为了证明本文模型的优越性，选择算法识别准确率、内存占有率和算法的稳健性作为评价指标。设实验样本识别正确的数目为 M，识别错误的数目为 N，则正确率为

$$A_{cc}=\frac{M}{M+N}\times 100\%。\tag{15}$$

对当前已有的算法进行复现，在相同环境和配置下进行实验，检测内存占有率。同时对视频进行加高斯白噪声处理，再把视频图像送入到本文网络模型并结合已有算法进行训练识别，图 6 所示为视频图像加噪前与加噪后的对比，图 6(a)为原始视频帧，图 6(b)为 $\sigma=0.2$(σ 为方差)时的噪声视频帧，图 6(c)为 $\sigma=0.4$ 时的噪声视频帧。

图 6 图片加入噪声后的前后对比。(a) 原始视频帧；(b) $\sigma=0.2$ 时的噪声视频帧；(c) $\sigma=0.4$ 时的噪声视频帧

Fig. 6 Comparison of video frames before and after adding noise to pictures. (a) Original video frames; (b) noise video frames with $\sigma=0.2$; (c) noise video frames with $\sigma=0.4$

3.4 实验结果对比与分析

经训练模型后对 20%的样本进行测试，实验表明在 Action Youtobe 数据集上，本文提出的网络模型的精度达到 94.38%，为了方便观测各行为之间的识别率，制作了数据集的混淆矩阵，对角线表示识别正确率，如表 3 所示。

由表 3 可知，有 8 种行为的精度达到了 90%以上，其中踢足球的识别中有 12.12%与高尔夫挥杆有所混淆，这是因为混淆行为之间具有较高的相似性，Inceptionv3 模型提取视频帧的深层特征有许多相同之处，Bi-LSTM 神经网络不能保证训练的模型权重在识别此类行为之间有明显差别，最后的注意力机制无法精确地感知其中的权重差异，从而导致识别率偏低。最后的遛狗是人与其他物体之间的互动行为，由于运动特征比较明显，神经网络在学习特征序列时无法分清两者的主次关系，所以容易与其他相似行为混淆。

为了说明本文算法的优越性，在同一数据集上分别用 LSTM、Bi-LSTM、Bi-LSTM-Attention 结合 Inceptionv3 模型进行实验，然后与现有的算法进行比较，如表 4 所示。

由表 4 可知，在 Action Youtobe 数据集中，本文提出的基于 Bi-LSTM-Attention 模型的人体行为识别算法在结合Inceptionv3模型后，可以得到优于

表 3 Action Youtobe 数据集行为识别混淆矩阵

Table 3 Action recognition confusion matrix of Action Youtobe dataset %

Category	Basketball	Biking	Diving	G-swing	H-riding	Soccer	Swing	Tennis	Jumping	Volleyball	Walking
Basketball	96.30	0	0	0	0	0	0	0	0	3.7	0
Biking	10.52	89.48	0	0	0	0	0	0	0	0	0
Diving	0	0	100.00	0	0	0	0	0	0	0	0
G-swing	0	0	0	96.67	0	0	3.33	0	0	0	0
H-riding	0	0	2.08	0	95.84	0	0	0	2.08	0	0
Soccer	0	0	0	12.12	0	87.88	0	0	0	0	0
Swing	0	0	0	0	0	0	96.55	0	0	3.45	0
Tennis	3.70	0	0	0	0	0	0	96.30	0	0	0
Jumping	4.35	0	0	0	0	0	0	0	95.65	0	0
Volleyball	0	0	9.52	0	0	0	0	0	0	90.48	0
Walking	0	0	0	3.84	0	3.84	0	0	0	3.84	88.48

表 4 Action Youtobe 数据集上本文算法与其他模型算法比较

Table 4 Comparison of proposed algorithm and other model algorithms on Action Youtobe dataset %

Algorithm	Accuracy	Memory occupancy	Accuracy (0.2)	Accuracy (0.4)
Binary CNN-Flow[18]	84.30	46	77.32	70.68
3D spatio-temporal[19]	88.00	—	—	—
Hierarchical clustering multi-task[7]	89.70	53	84.40	78.60
Deep-Temporal LSTM[15]	90.27	46	87.56	83.28
Discriminative representation[20]	91.60	—	—	—
Proposed DB-LSTM[16]	92.84	42	89.15	82.37
Fisher vectors[21]	93.80	—	—	—
Inceptionv3 + LSTM	89.53	31	83.54	76.54
Inceptionv3 + Bi-LSTM	92.81	33	88.38	82.82
Inceptionv3+ Bi-LSTM-Attention	94.38	37	92.56	89.24

Binary CNN-Flow、Discriminative representation、Proposed DB-LSTM、Deep-Temporal LSTM 4 种基于深度学习算法的精度，也优于另外 3 种基于手工特征提取的传统算法：Hierarchical clustering multi-task、Fisher vectors、3D spatio-temporal。本文在同一模型下对 LSTM、Bi-LSTM 两种算法进行了实验，实验结果表明，Bi-LSTM-Attention 模型使识别精度提升了 4.85% 和 1.57%。采用 Inceptionv3 模型提取视频图片特征，极大降低了网络参数，所提方法在内存占有率方面低于现有的算法，但是网络层次多于 LSTM、Bi-LSTM，所以在内存占有率方面略高于两种算法。同时，在数据集加上方差为 0.1 和 0.2 的噪声之后，本文算法的精度下降了 1.82% 和 5.14%，精度损失方面明显小于已有算法在视频加噪处理后的精度损失，证明了本文算法具有较高的稳健性。

本文将 Bi-LSTM-Attention 和普通深度学习方法以及传统手工特征提取算法对视频图像特征提取的结果进行可视化，绘制成的热力图如图 7 所示，颜色越深代表特征越明显。

如图 7 所示，传统手工方法容易忽略一些重要特征，对识别结果有着较大的影响。普通深度学习方法相比于传统的方法可以提取视频图像的深层特征，但是对一些重要的特征关注度较小，容易产生误

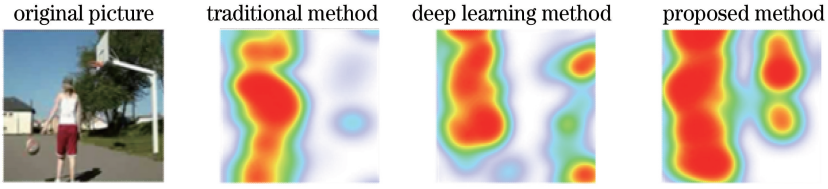

图 7 特征区域热力图

Fig. 7 Thermodynamic charts of feature regions

差。本文算法不仅可以充分提取视频图像的深层特征，还能重点关注对识别结果有较大影响的行为特征。因此本文算法相比于普通深度学习与传统手工特征提取方法有着较大的优势。

依然用 LSTM、Bi-LSTM、Bi-LSTM-Attention 方法结合 Inception v3 模型在 KTH 数据集上进行实验。首先选取前 120 个视频样本作为测试集，剩下的作为训练集，称其为 Dataset1，进行第一次交叉验证。以此类推每种方法一共进行 5 次交叉验证，取平均精度作为实验结果，如表 5 所示。

表 5 KTH 数据集交叉验证精度比较

Table 5 Accuracy comparison of cross validation for KTH dataset %

Algorithm	Dataset1	Dataset2	Dataset3	Dataset4	Dataset5	Average
Inception v3 ＋LSTM	97.50	82.50	97.50	86.67	87.50	90.33
Inception v3 ＋Bi-LSTM	99.17	87.50	100.00	93.33	93.33	94.67
Inception v3＋Bi-LSTM-attention	100.00	89.17	100.00	95.00	94.17	95.67

从表 5 可以看出，本文算法在 KTH 上的平均识别精度可以达到 95.67%，比 LSTM 和 Bi-LSTM 算法的结果分别高出了 5.34%和 1.00%。

表 6 为本文算法在 KTH 集上的混淆图，可以看出，击拳、挥手和拍手的识别率较高，但是挥手和拍手由于动作相似，它们之间存在 3%的误差。慢、快跑和行走三类行为之间的区分度不大，神经网络无法精确识别其中的相似特征，导致其中存在较大的误差。

表 6 KTH 数据集行为识别混淆矩阵

Table 6 Action recognition confusion matrix of KTH dataset

Action	Boxing	Handclapping	Handwaving	Jogging	Running	Walking
Boxing	99	0	0	0	0	1
Handclapping	0	97	3	0	0	0
Handwaving	0	3	97	0	0	0
Jogging	0	0	0	96	4	0
Running	0	0	0	5	93	2
Walking	0	0	0	4	4	92

表 7 为本文方法与现有方法的比较，可以看出，本文算法在 KTH 数据集上依然有良好的表现，同时也极大降低了内存占有率，在增加噪声后，本文算法精度损失也明显小于其他已有算法，证明了本文算法的可行性。

表 7 KTH 数据集上本文算法与其他模型算法比较

Table 7 Comparison of proposed algorithm and other model algorithms on KTH dataset %

Algorithm	Accuracy	Memory occupancy	Accuracy (0.2)	Accuracy (0.4)
3D CNN[11]	90.20	62	87.20	81.80
Spatio-temporal[6]	92.10	—	—	—
D-M and S-P feauters[22]	92.70	—	—	—
D-L slow feature[23]	93.10	58	0.80	85.40
Deep-Temporal LSTM[15]	93.90	46	90.10	84.60
CNN-LSTM[24]	94.20	—	—	—
Hierarchical clustering multi-task[7]	94.30	53	90.60	84.30
Inceptionv3 ＋ Bi-LSTM-Attention	95.67	37	93.80	90.27

4 结 论

提出了一种基于 Bi-LSTM-Attention 模型的人体行为识别算法。首先通过 Inceptionv3 模型提取视频帧中的深度特征信息，然后构建 Bi-LSTM 神经网络来学习模型中的时序特征信息，再通过注意力机制自适应感知网络权重，最后通过一层全连接层连接 Softmax 分类器对视频进行分类。实验结果表明，该算法能够在 Action Youtobe 和 KTH 数据集上达到高于现有方法的识别精度。与此前的方法相比，本文算法可以解决 LSTM 无法有效地提取视频帧之间的时序特征的问题，提升了识别率，但是本文算法在相似动作的识别上存在着不足，这将是以后研究的重点。

参 考 文 献

[1] Burić M, Pobar M, Kos M I. An overview of action recognition in videos[C]// 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), May 22-26, 2017, Opatija, Croatia. New York: IEEE, 2017: 1098-1103.

[2] Luo H L, Wang C J, Lu F. Survey of video behavior recognition[J]. Journal on Communications, 2018, 39(6): 169-180.
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.

[3] Willems G, Tuytelaars T, van Gool L. An efficient dense and scale-invariant spatio-temporal interest point detector[M]// Forsyth D, Torr P, Zisserman A. Computer vision-ECCV 2008. Lecture notes in computer science. Berlin, Heidelberg: Springer, 2008, 5303: 650-663.

[4] Rapantzikos K, Avrithis Y, Kollias S. Dense saliency-based spatiotemporal feature points for action recognition[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 1454-1461.

[5] Abdulmunem A, Lai Y K, Sun X F. Saliency guided local and global descriptors for effective action recognition[J]. Computational Visual Media, 2016, 2(1): 97-106.

[6] Luo J J, Wang W, Qi H R. Spatio-temporal feature extraction and representation for RGB-D human action recognition[J]. Pattern Recognition Letters, 2014, 50: 139-148.

[7] Liu A A, Su Y T, Nie W Z, *et al*. Hierarchical clustering multi-task learning for joint human action grouping and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1): 102-114.

[8] Liu Z, Huang J T, Feng X. Action recognition model construction based on multi-scale deep convolution neural network [J]. Optics and Precision Engineering, 2017, 25(3): 799-805.
刘智, 黄江涛, 冯欣. 构建多尺度深度卷积神经网络行为识别模型[J]. 光学 精密工程, 2017, 25(3): 799-805.

[9] Zhu Y, Zhao J K, Wang Y N, *et al*. A review of human action recognition based on deep learning[J]. Acta Automatica Sinica, 2016, 42(6): 848-857.
朱煜, 赵江坤, 王逸宁, 等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848-857.

[10] Charalampous K, Gasteratos A. On-line deep learning method for action recognition[J]. Pattern Analysis and Applications, 2016, 19(2): 337-354.

[11] Ji S W, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[12] Donahue J, Hendricks L A, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 2625-2634.

[13] Gammulle H, Denman S, Sridharan S, *et al*. Two stream LSTM: a deep fusion framework for human action recognition[C]// 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), March 24-31, 2017, Santa Rosa, CA, USA. New York: IEEE, 2017: 177-186.

[14] Li Q H, Li A H, Wang T, *et al*. Double-stream convolutional networks with sequential optical flow image for action recognition[J]. Acta Optica Sinica, 2018, 38(6): 0615002.
李庆辉, 李艾华, 王涛, 等. 结合有序光流图和双流卷积网络的行为识别[J]. 光学学报, 2018, 38(6): 0615002.

[15] Das S, Koperski M, Bremond F, *et al*. Deep-temporal LSTM for daily living action recognition[C]// 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), November 27-30, 2018, Auckland, New Zealand. New York: IEEE, 2018: 18455900.

[16] Ullah A, Ahmad J, Muhammad K, *et al*. Action recognition in video sequences using deep bi-directional LSTM with CNN features[J]. IEEE Access, 2018, 6: 1155-1166.

[17] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 2818-2826.

[18] Ravanbakhsh M, Mousavi H, Rastegari M, *et al*. Action recognition with image based CNN features[J/

OL]. (2015-12-13)[2019-01-02]. https://arxiv.org/abs/1512.03980.

[19] Yang X D, Tian Y L. Action recognition using super sparse coding vector with spatio-temporal awareness[M]// Fleet D, Pajdla T, Schiele B, *et al*. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8690: 727-741.

[20] Wang J, Liu Z C, Wu Y, *et al*. Mining actionlet ensemble for action recognition with depth cameras[C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 16-21, 2012, Providence, RI, USA. New York: IEEE, 2012: 1290-1297.

[21] Peng X J, Zou C Q, Qiao Y, *et al*. Action recognition with stacked fisher vectors[M]// Fleet D, Pajdla T, Schiele B, *et al*. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 581-595.

[22] Li Y D, Xu X P. Human action recognition by decision-making level fusion based on spatial-temporal features[J]. Acta Optica Sinica, 2018, 38(8): 0810001.
李艳荻, 徐熙平. 基于空-时域特征决策级融合的人体行为识别算法[J]. 光学学报, 2018, 38(8): 0810001.

[23] Sun L, Jia K, Chan T H, *et al*. DL-SFA: deeply-learned slow feature analysis for action recognition[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 2625-2632.

[24] Huang Y W, Wan C L, Feng H. Multi-feature fusion human behavior recognition algorithm based on convolutional neural network and long short term memory neural network[J]. Laser & Optoelectronics Progress, 2019, 56(7): 071505.
黄友文, 万超伦, 冯恒. 基于卷积神经网络与长短期记忆神经网络的多特征融合人体行为识别算法[J]. 激光与光电子学进展, 2019, 56(7): 071505.