文章编号：2095-2163(2020)10-0086-07 中图分类号：TP391.41 文献标志码：A

基于多流 3D 融合网络的人体行为识别

戎 炜，张天雨
（合肥工业大学 计算机与信息学院，合肥 230601）

摘 要：人体行为识别是当前计算机视觉领域的一个重要研究分支。针对视频人体行为识别任务需要大型数据集预训练以及无法有效利用跨时间信息的问题，本文提出了基于双流卷积网络与膨胀 3D 卷积网络的深度神经网络模型，并重新设计网络结构，命名为多流 3D 融合网络。首先，利用改进的双流网络与膨胀 3D 网络提取人物动作特征；其次，利用分段长短期记忆网络提取时间特征；最后，利用残差连接方法融合特征，得到最终的个体识别结果，实现了精确的个体行为识别。在 volleyball 数据集上的实验结果表明，本文提出的方法优于当前的一些先进方法。

关键词：行为识别；膨胀 3D 卷积网络；双流卷积网络

Human action recognition based on multi-stream 3D fusion network

RONG Wei, ZHANG Tianyu
(College of Computer and Information, Hefei University of Technology, Hefei 230601, China)

【Abstract】 Human action recognition is a significant research branch in modern computer vision. Aiming at the problem that the current benchmark requires pre-training of large datasets and the inability to effectively use cross-time information, a deep neural network model based on two-stream convolutional network and inflated-3D convolutional network is proposed, and the network structure is redesigned, named multi-stream 3D fusion network. First, the improved two-stream network and inflated-3D network are used to extract the motion features of individuals. Then, the segmented long-short-term memory network extracts the time features, Finally, the residual connection operation fuses the features to obtain the final individual recognition results, which is proved to achieve accurate individual activity recognition. Results on volleyball dataset demonstrate that our approach significantly outperforms state-of-the-art techniques.

【Key words】 action recognition; inflated-3D convolutional network; two-stream convolutional network

0 引 言

行为识别是指从视频帧序列中提取出与目标行为相关的有用信息，并采用合适的方式进行数据表达，通过解释这些行为视觉信息，达到对人的行为模式分析和识别的目的。行为识别按识别对象区分为个体行为识别与群组行为识别。研究者针对个体行为识别任务，提出了一些方法。受到在静态图像上成功使用卷积神经网络的鼓舞，许多研究人员开发了用于视频理解和行为识别的方法[1]。最近的大多数作品都受到 Simonyan 等人提出的双流卷积神经网络的启发，其中合并了从 RGB 图像和光流图像中提取的空间与时间信息；另一方面，对于视频行为识别，3D 卷积网络亦是该领域的重要研究热点，已被广泛应用于行为识别任务中。但是，3D 卷积网络的预训练过程不仅需要大量的视频数据，而且还需要大量的硬件资源。

对于行为识别中双流网络与 3D 卷积网络各自的局限性，本文提出了一个新的网络模型。该网络模型将双流网络与 3D 卷积网络的特性结合，同时重新设计网络结构，以弥补两种网络的缺陷，使得该网络模型能出色地完成个体行为识别任务。

1 多流 3D 融合网络

本文提出的改良多流 3D 融合网络（Multi-stream 3D Fusion Network，M3DFN）模型将双流网络和 3D 卷积网络的特性结合，并加以改良，以提升在群体中个体行为识别的性能。多流 3D 融合网络的结构如图 1 所示。网络由输入采样模块，目标定位提取模块，多流 3D 卷积模块，分类 LSTM 模块等主要部分组成。在视频行为识别任务中，对于输入视频序列，采用时序分割方法，将视频序列分为若干帧一个片段，在每个片段中随机采样一帧图像，之后使用 Faster R-CNN 网络对图像中的人物进行目标定位。取人物帧图像前后若干帧组成图像序列，对其进行光流提取后得到光流信息特征图；将输入的图

作者简介：戎 炜（1994-），男，硕士研究生，主要研究方向：计算机视觉；张天雨（1996-），男，硕士研究生，主要研究方向：计算机视觉。
通信作者：戎 炜 Email：550401736@qq.com

收稿日期：2020-04-27

像分为多帧图像序列，采样得到的单帧图像以及多帧光流图序列，输入多流 3D 卷积模块中，输出的个体特征进行特征连接操作得到全局特征；将各个多流 3D 卷积模块输出的个体特征输入到分段 LSTM 模块中，输出的融合特征再次与全局特征融合；最后，经由全连接层与 softmax 分类操作得到最终个体行为识别结果。

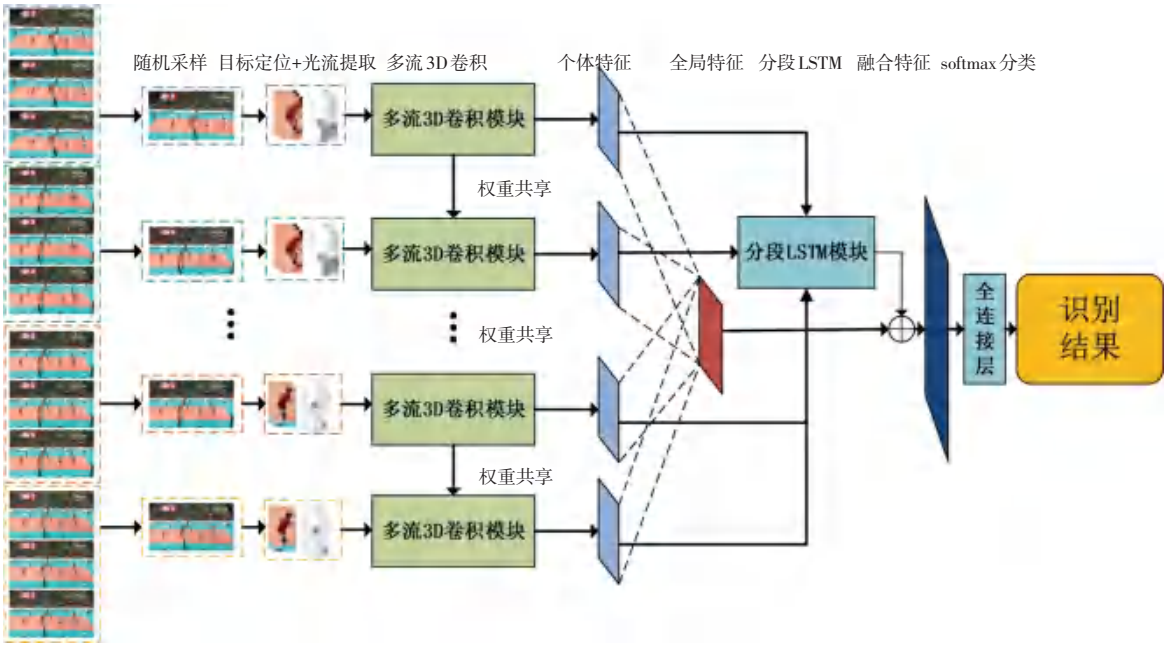


图 1　多流 3D 融合网络结构图

Fig. 1　The structure of multi-stream 3D fusion network

2　视频特征的多流提取与处理

2.1　膨胀 3D 卷积网络

多流 3D 卷积模块的结构如图 2 所示。双流网络部分沿用经典的时序分割网络，而膨胀 3D 卷积网络部分则是在已有的 2D ResNet101 上进行 2D 膨胀操作，将其扩充为 3D 卷积网络。膨胀 3D 卷积网络的输入是以随机采样得到的单帧 RGB 图像为中心的 RGB 图像序列，输出则是个体特征。为了与 2D 卷积输出的特征维度匹配，膨胀 3D 卷积网络的输出特征将会被压缩为 2D 尺寸。

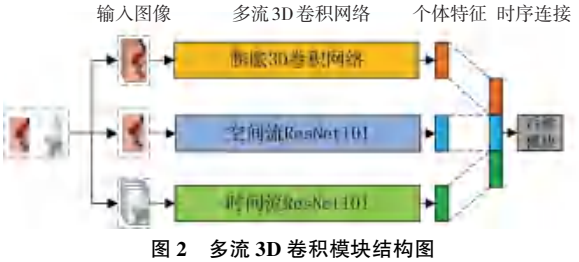


图 2　多流 3D 卷积模块结构图

Fig. 2　The structure of multi-stream 3D convolutional module

残差神经网络（Residual neural Network，ResNet）最早是由 He 等人提出的，并在图像识别任务有出色表现[2]。代替直接拟合的基础映射 $H(x)$，残差网络将原始映射改良为 $F(x)+x$ 以拟合残差映射。这类研究表明了这种连接之前输入的方式十分有效[3]。残差连接方式可以表示为式(1)和式(2)：

$$y_i = x_i + F(x_i, W), \tag{1}$$

$$x_{i+1} = f(y_i). \tag{2}$$

其中，x_i 与 x_{i+1} 代表第 i 层的输入和输出；$F(x_i, W)$ 代表残差映射；$f()$ 代表 ReLU 过滤函数。对于多于 50 层的网络，残差映射 $F(x_i, W)$ 则是由 3 层一组的形式组成。

由于 3D 卷积网络在空间卷积结构上与对应的 2D 卷积很相似。因此可以将 ImageNet 预先徐连的 2D 参数视为 3D 内核的一部分。可以沿时间维度将 2D 参数直接复制到 3D 内核中，这就是 2D 膨胀操作。但是由于参数不足以支撑起多出来的时间维度，仍然需要重新设计时间结构。受到 I3D 网络提出的扩展操作的启发，本文采用 2D 膨胀操作，用于引导 ImageNet 预训练参数。具体的思想是用 3 个 2D 卷积核来组成一个 3D 卷积核，这些 2D 卷积核是从对应的 ImageNet 预训练的 2D 卷积层的同一通道中复制的。于是参数的尺寸可以由正方形转换为立方体。这些操作可以描述为公式(3)和公式(4)：

$$K_m^l = \{k_m^l, k_m^l, k_m^l\}, \tag{3}$$

$$K^l = C_l \sum_{i=0}^{m-1} K_i^l, \tag{4}$$

其中，k_m^l 代表第 l 层的第 m 通道的预训练 2D 卷积核，K_m^l 则表示对应的 3D 卷积核。C_l 代表将所有卷积核融合成一个卷积核 K^l 的操作。

本文提出的多流 3D 卷积模块中的膨胀 3D 卷积网络是由 2D 的 ResNet101 经由 2D 膨胀操作变化得到的。具体操作如图 3 所示：将输入卷积的大小由 7×7 卷积变为 3×7×7 卷积。padding 的尺寸由 3×3 变为 1×3×3。3×3 卷积包括最大池化卷积变为 3×3×3 卷积，1×1 卷积变为 1×1×1 卷积。时间维度的步长均设为 1，空间维度的步长保持不变，最大池化卷积在的时间与空间维度的步长也都保持不变。膨胀 3D 卷积网络的预训练参数由 ImageNet 预训练的对应 2D 卷积网络提供，因此不需要如 Kinetics 之类的数据集预训练，节省了大量的时间与计算开销。

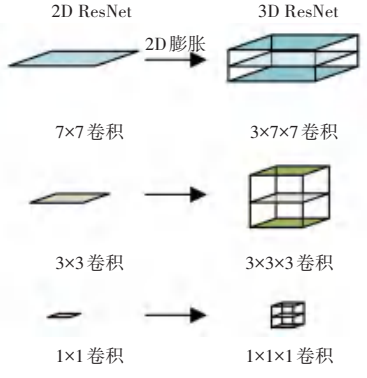


图 3 2D 膨胀操作示意图

Fig. 3 The 2D-inflated operation

2.2 双流卷积网络

双流卷积网络由两个独立的空间流卷积网络和时间流网络构成。空间流网络将 RGB 图像作为输入，而时间流网络则使用堆叠的光流图像作为输入。大量文献表明，较深的卷积网络可以提高双流网络的整体性能。特别是 VGG-16，GoogleNet 和 BN-Inception 在空间流和时间流上的性能都得到了验证。但 ResNet101 展示了其捕获静态图像特征的能力，因此选用 ResNet101 作为空间流和时间流的基准网络。空间流输入方面，采用单帧 RGB 图像已被证实十分有效。时间流输入方面，采用标准的 10 帧连续光流图像序列。Feichtenhofer 等人的实验证明了融合特征的重要性，后期融合特征可以达到最佳融合，而早期的融合虽然需要的参数较后期少，但达到的性能不如后期融合。因此，本文采用最后一层融合特征的方式构造双流网络。在最后的融合中，本文采用了特征串联的方式，不仅串联时间流和空间流，也让膨胀 3D 卷积网络的输出特征参与串联过程。串联融合得到的特征成为分段 LSTM 的输入。双流 ResNet101 由 ImageNet 网络预训练，鉴于膨胀 3D 卷积网络的特性，可以从双流网络中共享参数。因此只需预训练双流网络，便可为膨胀 3D 卷积网络提供参数。

3 时间信息处理与特征融合

3.1 分段 LSTM 网络

视频内每个图像帧的变化都可能包含其他信息，这些信息可能对确定整个视频的人体行为有所贡献。最能直接提取并利用这些信息的模型之一是循环神经网络（RNN）。RNN 可以通过隐藏状态单元设计学习时间动态信息。但是由于 RNN 存在的长时依赖问题，使用 LSTM 代替 RNN 是较好的选择。然而，更深的 LSTM 层不一定有助于获得更好的动作识别性能，因为之前的双流卷积网络与 3D 卷积网络已经提供了足够强大的学习性能。

本文采用 LSTM 单元与时间池化层的结合来提取时间动态信息，构造分段 LSTM 网络，如图 4 所示。输入特征为串联的 3 种特征序列。经过与采样阶段相同数量的分段后，经过 BN（Batch Normalization）操作后，使用时间池化层从每个片段中提取区别特征，再输入 LSTM 中提取嵌入特征。时间池化层可以是平均池化层或者最大池化层，本文选用最大池化层。时间池化层从 3D，空间和时间流串联的特征向量中提取区别特征。而 LSTM 将提取整个视频的嵌入特征。其本质上是学习非线性特征组合及其随时间变化的分段表示的机制。

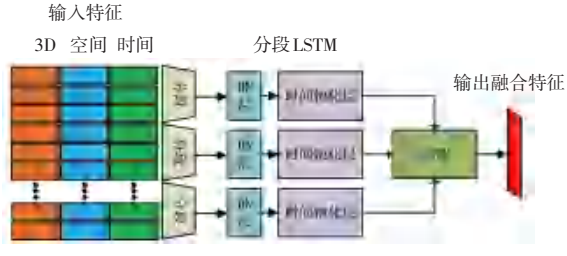


图 4 分段 LSTM 结构图

Fig. 4 The structure of segment LSTM

3.2 分支特征融合

经过各个模块的提取输出的分支特征需要按照顺序进行汇总。本文评估了 3 种特征融合方法，如图 5 所示。

最简单直接的方法便是图 5 中左侧的直接连接。将每个分支特征向量按时间顺序连接到组合的特征向量中，之后直接输入全连层和分类层。聚合信息的第二种方法是图 5 中间的操作，添加了全连接层和 Dropout 操作，全连接层能进一步处理组合

的特征向量，进一步提高识别的准确率。图 5 中右侧的是基于第二种方法的第 3 种方法残差连接，添加了残差全连接层，从而聚合从视频中提取的特征。本文采用第三种方法，公式(5)和公式(6)为

$$x_c = \{x_0, x_1, \ldots, x_{n-1}\}, \tag{5}$$

$$x_t = H_c(W_c X_c + x_c). \tag{6}$$

其中，x_c 为各个分支特征 x_n 合并而成，并进行残差连接操作，W_c 代表残差连接层的权重，H_c 代表 ReLU 函数与 Dropout 的结合操作。实验结果表明，组合特征向量的残差连接处理是有益的，连接输入特征向量丰富了特征中的信息，提升了识别性能。

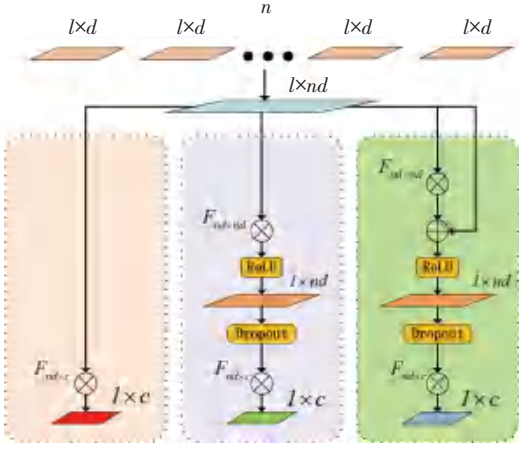


图 5　3 种特征融合方法示意图

Fig. 5　Three feature fusion operations

4　实验结果及分析

4.1　数据集介绍

为了证明本文提出的网络模型能有效完成个体行为识别任务，本文在 volleyball 数据集上对模型进行验证。该数据集是用于群体行为识别的数据集，但因为对场景中的每个个体的动作及位置都设置了标签，因此也适用于个体行为识别。Volleyball 数据集的视频均为排球比赛，收集自 YouTube 视频网站。该数据集包含了 55 场排球比赛的实况录像，并且制作者为其中的 4830 帧制作了位置与行为标注。每个运动员个体都以一个边界框的坐标和 9 种个体动作之一进行标注，而这 9 种个体动作对应 8 种群体行为，表明在场景中的某个群组发生的群组行为类别。本文将 volleyball 数据集的 2/3 用于训练，1/3 用于测试。

4.2　评价指标

文中选择准确率(Accuracy)指标来评价方法和模型的性能。准确率是群组行为识别任务广泛采用的指标，准确率计算方法如式(7)：

$$Acc = \left(\sum_{j=1}^{N} n_{jj}\right) / \left(\sum_{i=1}^{N}\sum_{j=1}^{N} n_{ij}\right). \tag{7}$$

其中，n_{ij} 指真实标签是 i，分类预测结果标签是 j 的样本数量。n_{jj} 是 n_{ij} 的特殊情况，代表真实标签和分类结果标签均为 j。N 代表参与测试和评价的样本总数量，Acc 代表准确率。准确率越高代表方法和模型的效果越好。

4.3　实验环境

本文在 64 位系统 Ubuntu16.04 上安装了 pytorch 深度学习框架，该计算机 GPU 由两块 NVIDIA GeForce GTX 1080 与一块 NVIDIA GeForce TITAN xp 组成，共有四块 GPU。CPU 采用 Intel Core i7-8700k 型号。内存大小为 48G，编程环境选择 python3.6 环境。

4.4　实验方法

本文的实验方法选择标注帧的前五帧与后四帧，包括标注帧在内的 10 帧时序连续图像作为输入。在消融实验中，将调整包括标注帧在内的时序连续图像的数量，比如调整为 25 帧时序连续图像。本文中的卷积神经网络采用残差网络与密集网络的 3D 膨胀版本，该网络的特性是不需要预训练也能表现出较好的性能，省去了庞大的预训练开销。输入图像需统一调整分辨率为 224×224，并经过数据扩充处理。本文的数据扩充方法为多尺度随即裁切，即裁切由最小长度与尺度乘积定义的区域，比例从 1.0，0.875，0.75，0.66 中随机选择。同时，对每 3 帧图像，执行水平翻转操作的概率为 50%。之后分别提取裁剪视频帧的外观特征与运动特征以满足时序分割部分输入的需要。本文的 LSTM 部分采用单层 LSTM 网络，输入的特征向量为 4096 维，LSTM 隐藏单元为 512 个。

本文实验采用 Faster R-CNN 网络作为目标检测方法，并对检测出的场景中的个体目标提取外观信息和运动信息，并送入空间流与时间流网络。在空间流与时间流网络中，经过膨胀 3D 卷积层提取融合操作，输出的特征经过连接操作，进入 LSTM 网络提取跨时间信息，并得到个体行为的特征表达，经过 softmax 层分类，作出最终的行为预测结果。同时本文也将使用真实位置标注的实验结果作为对比。

本文网络模型的优化算法选用 Adam 优化算法，Dropout 参数的值设置为 0.5，以防止过拟合现象。模型的初始学习率设置为 0.001，衰减设置为每

个周期的学习率衰减为上个周期的 0.75。这是因为传统的梯度下降策略将导致损失的持续增长,并且过快的梯度更新更容易过拟合。批处理数据大小为 128,即网络每个周期处理 128 段视频序列。训练周期为 340 个周期,即网络对整个数据集训练 340 次。

4.5 对比实验结果分析

为了研究本文提出的多流 3D 融合网络在个体行为识别任务中的提升效果,本文在 volleyball 数据集上进行了模块递进的消融实验,各方法消融实验结果见表 1。

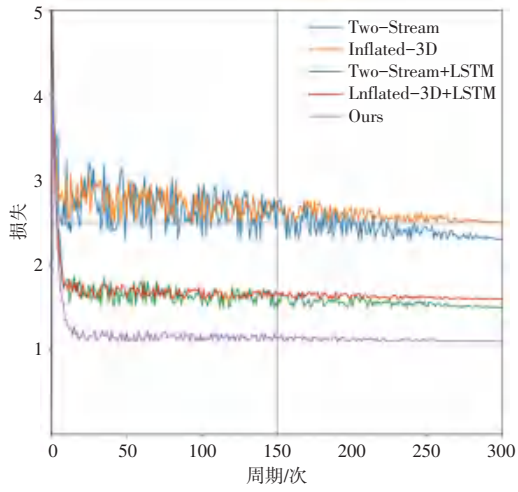
表 1 Volleyball 数据集上的消融实验结果

Tab. 1 Ablation results onvolleyball dataset

实验方法	准确率/%
Two-Stream	71.2
Inflated-3D	68.8
Two-Stream+LSTM	73.7
Inflated-3D+LSTM	70.8
Ours	75.4

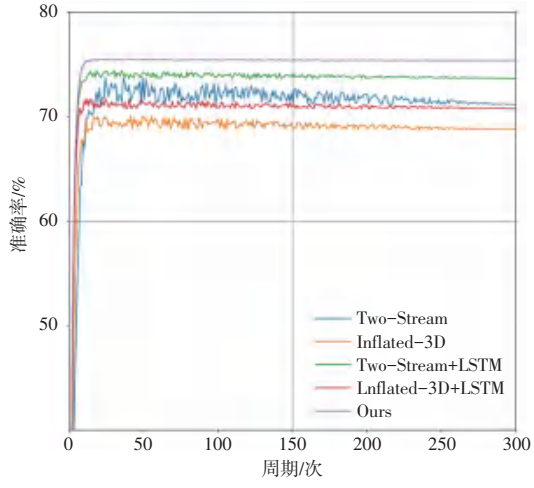
在表 1 中,将本文方法与各方法进行了对比,该对比实验未使用真实位置标签,而是利用 Faster-RCNN 网络对群组中个体进行空间定位,同时对输入视频段的采样设置为 3 帧一段。表 1 中 Two-Stream 表示传统时序分割双流网络方法;Inflated-3D 表示经过膨胀操作后的 3DresNet101 网络方法;Two_Stream+LSTM 与 Inflated+LSTM 代表为这两种网络添加 LSTM 层后形成的网络方法;Ours 代表本文提出的方法。由表 1 可知,双流网络的识别准确率要高于膨胀 3D 卷积网络,而在加入 LSTM 层后,这两种方法的识别准确率也都有所上升,不过双流网络方法的识别准确率依然要高于膨胀 3D 卷积方法的识别准确率。而本文提出的方法由于融合了这几种网络的特点,其识别准确率均高于这几种网络的识别准确率。

表 1 中的方法的损失与准确率收敛曲线如图 6 所示。图 6(a)中,双流网络的时间流需要对光流图提取时间信息,导致整体训练速度较慢,损失震荡较明显,需要较长时间收敛。而膨胀 3D 网络将时间信息作为一个维度的信息进行提取,训练速度较双流网络要快,损失对比双流网络收敛较快,但最终收敛损失比双流网络要高。添加分段 LSTM 模块的双流网络与膨胀 3D 网络损失收敛更快,这是 LSTM 网络更好地提取时间信息的缘故。本文提出的方法结合了多种网络的优点,损失下降最快,且震荡较小,最终收敛损失也最低。图 6(b)中也能看出,本文提出的方法的准确率最高,对比双流 LSTM 网络与膨胀 3D 网络的准确率,分别提高了 1.7%和 4.6%。



(a) 损失收敛曲线

(a)



(b) 准确率收敛曲线

(b)

图 6 不同方法的损失与准确率收敛曲线

Fig. 6 The loss and accuracy curves of different methods

不同融合方法的对比实验结果见表 2。3 种融合方法在 volleyball 数据集上的实验准确率如图 7 所示,Direct-Connection 代表直接连接融合,Fully-Connection 代表全连接融合,Residual-Connection 代表基于全连接融合的残差连接融合。其中残差连接融合准确率最高,对比直接连接融合与全连接融合分别提升了 2.6%和 1.9%。实验证明,使用残差连接融合处理组合特征向量丰富了特征中的信息,提升了识别性能。

表 2　不同融合方法的对比实验结果

Tab. 2　Results of different fusion operation

实验方法	准确率/%
Direct-Connection	72.8
Fully-Connection	73.5
Residual-Connection	75.4

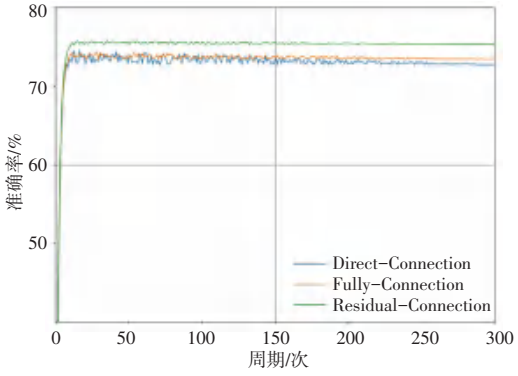


图 7　3 种融合方法的准确率收敛曲线

Fig. 7　The accuracy curves of three fusion operations

4.6　与其他方法对比分析

将本文提出的方法与 Bagautdinov 提出的方法进行了对比，实验结果见表 3。

表 3　Volleyball 数据集上的对比实验结果

Tab. 3　Results on volleyball dataset

实验方法	准确率/%
Bagautdinov-single	77.8
Bagautdinov-temporal	77.9
Ours-3S	75.4
Ours-5S	76.3
Ours-GT-3S	81.2
Ours-GT-5S	82.8

Bagautdinov-single 代表输入图像帧数为 1 帧；Bagautdinov-temporal 代表输入图像帧数为 10 帧序列；Ours 代表本文提出的方法，3S 代表输入视频分割方法为 3 帧一段，5S 则代表以 5 帧一段进行分割，GT 代表使用真实位置标签进行目标定位，否则代表使用 Faster R-CNN 网络进行目标定位。由表 3 可知，本文提出的方法在不使用真实位置标注的情况下，识别准确率要低于 Bagautdinov 提出的方法。而使用了真实位置标注后，本文的方法的识别准确率则高于 Bagautdinov 方法。另外，5 帧分割方法的识别准确率要高于 3 帧分割方法，这是因为获得了更多的输入帧，从输入中提取的时间信息更加丰富。而对于可能出现的模糊、遮挡情况，Faster R-CNN 网络检测极易出现偏差，且对于某些实际边界框个数少于真实边界框标签个数的场景，本文采用将其特征置 0 的处理方式，这同样也会影响个体行为识别结果，而真实位置标注也不会存在这样的问题。

4.7　混淆矩阵分析

使用 Faster R-CNN 定位与真实位置标注定位的实验混淆矩阵如图 8 与图 9 所示。可以看出，使用真实位置标注定位的准确率要高于使用 Faster R-CNN定位的准确率。两者在 setting，jumping，moving 等动作的识别准确率上有较大差异。这是因为 Faster R-CNN 定位的目标位置与真实位置有偏差，以及忽略某些人物的位置预测所造成的。如图 10 所示，蓝框为 Faster R-CNN 检测到的人物位置框，黄圈中则为未检测的人物。由于场景与人物互相遮挡，以及实际边界框个数与标签不匹配等问题，会造成人物定位的偏差与遗漏，从而导致识别准确率下降。

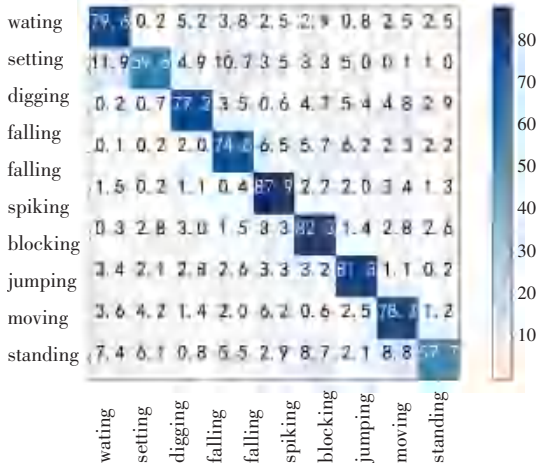


图 8　Faster R-CNN 定位的混淆矩阵

Fig. 8　Confusion matrix with Faster R-CNN

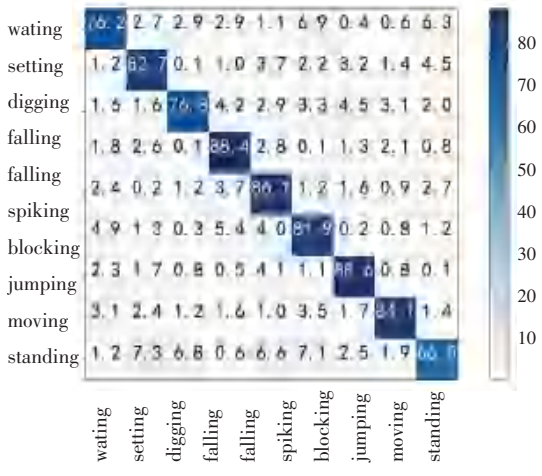


图 9　真实位置标注的混淆矩阵

Fig. 9　Confusion matrix with groundtruth label

（下转第 94 页）