文章编号：2095-2163(2020)06-0031-06 中图分类号：TP391 文献标志码：A

基于时空信息融合的时序动作定位

王 倩，范冬艳
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：时序动作定位任务需要识别出一段长视频中的动作类别以及动作的起止时间，候选区域的选择是影响到识别效果和效率的重要因素。提出一种基于时空特征融合的候选区域提取网络，充分利用视频分割段的时间特征和空间特征来判断是否为候选区域。接着将候选区域输入到训练的 CDC 网络中进行帧级粒度上的动作分类。最后训练动作状态检测网络，对得到的候选区域进行修补，从而可以得到更为精确的动作发生的时间区域。在 THUMOS'14 数据集上进行实验，结果证明该方法可以有效地进行未剪辑视频的时序动作定位，相对现有方法达到了较高的精度。

关键词：时序动作定位；时空特征；候选区域；CDC 网络；动作状态检测网络

Temporal action location based on spatio-temporal information fusion

WANG Qian, FAN Dongyan
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 The temporal action location task needs to identify the action classes, the start and end time of actions in a long video. The selection of candidate areas is an important factor affecting the recognition effect and efficiency. In this paper, a candidate region extraction network based on fusion of spatio-temporal features is proposed, which makes full use of the temporal and spatial features of video segmentation segment to determine whether it is a candidate region. Then, the candidate regions are input into the trained CDC network for frame level granularity action classification. Finally, the action state detection network is trained to repair the candidate regions, so as to get a more accurate time region of action occurrence. Experiments on THUMOS'14 dataset show that the proposed method can be effectively applied to temporal action localization in untrimmed videos, and has a higher accuracy than the existing methods.

【Key words】 temporal action localization; temporal and spatial features; candidate regions; CDC network; action state detection network

0 引 言

随着计算机视觉等相关技术的发展，深度学习已经在视频动作识别领域[1-5]取得了巨大成果。但是实际应用中的视频通常是不受约束的，包含多个动作实例和背景场景或其他活动的视频内容。时序动作定位是一个重要而又具有挑战性的问题，给定一个包含多个动作实例和背景内容的长而未修剪的视频，不仅需要识别它们的动作类别，还需要定位每个实例的起始时间和结束时间。

许多先进的系统使用段级分类器来选择和排列预先设定边界的建议段[6-9]。然而，一个理想的模型应该超越节段的级别，在细粒度的时间尺度上进行密集的预测以确定精确的时间边界。例如 S-CNN[6]的 C3D 卷积神经网络[3]从 conv1a 到 conv5b 层，将输入视频的时间长度减少了 8 倍。为此，Zheng Shou 等设计了一个新的 convolutional-de-convolutional(CDC)网络[10]，该网络的顶部是一个 C3D 卷积神经网络用于提取视频的时空特征信息，用来判断动作类型。然后采用 CDC 过滤器同时执行所需的时间向上采样和空间下采样操作，以在帧级粒度上预测动作。CDC 网络不仅在每一帧检测行为上实现性能优越，同时也显著提高了时间边界的精确性。但是 CDC 网络在候选区域的选取算法和时间边界的定位上还有待提高，主要有两个问题：(1)CDC 进行时序动作定位预测的输入为原始视频的候选片段，候选片段的选择会影响到时序动作识别的效果和效率，若识别不准确，不仅影响识别结果的准确率，还会耗费时间去识别不准确的候选区域。S-CNN 建议片段选择算法将密集间隔采样的 RGB 图输入 C3D 卷积神经网络进行预测，没有充分利用

基金项目：国家自然科学基金青年基金项目(61802251)；上海市科学技术委员会科研计划项目(16dz1206000)；面向广义宽基线立体像对的目标三维重建技术研究(61801286)；上海工程技术大学科研项目(E3-0903-19-01053)。

作者简介：王 倩(1995-)，女，硕士研究生，主要研究方向：计算机视觉与图形处理；范冬艳(1995-)，女，硕士研究生，主要研究方向：计算机视觉与图形处理。

收稿日期：2020-03-23

视频的时空特征。(2)结合候选区域与帧级分数并利用阈值定位边界点可以得到最终的时序动作检测结果。然而,检测得到的动作的起始和终止坐标与真值之间还有着较大的偏移。基于以上两个问题,本文提出了时空特征融合时序动作定位模型(spatio-temporal feature fusion temporal action localization model,STFF-CDC)。

针对问题 1,为了充分融合视频中的时空信息,并以相当低的成本保存相关信息,文献[11]在 TSN 网络的基础上,结合 C3D 卷积神经网络,提出了时空特征融合动作识别模型。该模型能够充分利用视频的时空特征,有效且高效的识别视频的动作类型。将该模型用于候选区域的选择网络,可以充分融合视频的时空特征,提高候选区域提取的准确率。

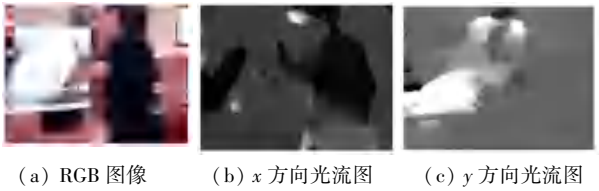
针对问题 2,为了解决定位的动作起始和终止坐标与真实值存在较大偏移的问题,本文提出了一个动作起始终止状态判断网络。将检测结果的时间区域扩大,之后再用起始点和结束点周围的帧为数据集训练 DenseNet 网络[12],利用训练好的模型判断起始帧和结束帧,从而重新定位得到更加精确动作起始边界。

1 相关工作

1.1 RGB 图和光流图

视频的 RGB 图像使 3 个颜色通道(红色 R,绿色 G 和蓝色 B 来存储像素信息,这些像素信息包含了视频的外形信息,如图 1(a)所示。由于视频的动作识别与视频中某些对象密切相关,外形信息是动作识别的重要信息,因此 RGB 图像可以提取视频的空间特征。

光流场是指图像中所有像素点构成的一种二维瞬时速度场,其中的二维速度矢量是景物中可见点的三维速度矢量在成像表面的投影。所以光流不仅包含了被观察物体的运动信息,而且还包含有关景物三维结构的丰富信息[13]。视频的光流图包含了视频的运动信息,如图 1(b)、1(c)所示,分别为水平方向(x 方向)和垂直方向(y 方向)光流图样例。



(a) RGB 图像 (b) x 方向光流图 (c) y 方向光流图

(a) RGB image (b) x - direction optical flow image (c) y - direction optical flow image

图 1 披萨抛掷类的 RGB 和光流图示例

Fig. 1 Examples of RGB and optical flow images for pizza thrower classes

1.2 时序动作检测

时序动作检测任务的目的是识别一段未剪辑长视频中的动作类别以及动作的起止时间。近年来,出现一些方法用于时序动作定位任务。S-CNN 是较为典型的一种方法,S-CNN 框架主要分为多尺度段的生成、Segment-CNN、后处理 3 个部分,Segment-CNN 包括建议网络,分类网络和定位网络 3 个子网络,均使用了 C3D 卷积神经网络。建议网络是提取候选区域的网络,它的输出为两类,即预测该片段是动作的概率及是背景的概率;分类网络的作用是尝试做一个用来识别动作的种类的分类网络,为定位网络做初始化;定位网络的输出为 $K+1$ 个类别(包括背景类)的分数,这个分数是该 segment 是某类动作的置信度分数。后处理是在测试阶段进行的,具体方法是对定位网络的输出分数进行非极大化抑制(NMS)来移除重叠,对于时序上重叠的动作,通过 NMS 去除分数低的,保留分数高的。CDC 使用的候选区域选择算法是 S-CNN 建议网络的候选区域提取算法,即改进 C3D 卷积神经网络的 SoftMax 层进行候选区域的判定。

2 基于时空信息融合的时序动作检测网络

2.1 网络整体框架

改进的 CDC 卷积神经网络训练时的输入为带有帧级标签的窗口训练集,预测时的输入为经过基于特征融合的候选片段生成算法生成的建议片段。接着经过 3D 卷积神经网络进行语义特征提取,这时视频的时间维度会缩小为 $L/8$。为了恢复时间维度上的信息,来进行帧级粒度上的动作识别,再经过 3D 反卷积神经网络来对视频的时间维度进行上采样,空间维度进行下采样。输出每帧的动作类别分数,最后训练动作状态检测网络进行时间边界的调整。框架流程见下图 2 所示。

2.2 动作候选区域提取算法

2.2.1 S-CNN 候选区域提取算法

S-CNN 候选区域提取算法使用 Du Tran 等[3]提出的 C3D 卷积神经网络作为动作分类网络,C3D 卷积神经网络与 2D 卷积神经网络不同,它有选择地兼顾运动和外观。在跳高的例子中,特征先是集中在整个人身上,然后跟踪其余帧上的人体跳高的动作。同样在化唇妆例子中,它首先聚焦在嘴唇上,然后在化妆时跟踪嘴唇周围发生的动作。C3D 卷积神经网络不仅对空间的的水平和竖直维度进行卷积,对时间维度也进行了卷积,以更好地提取时间和空间特征,保持时空特征的相关性。

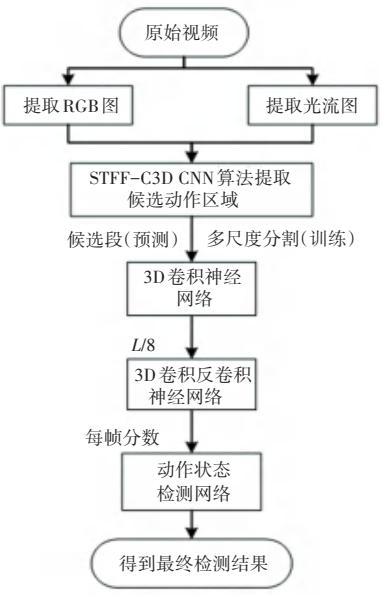


图 2　时空融合时序动作检测网络框架图

Fig. 2　Spatio-temporal feature fusion temporal action detection network frame diagram

S-CNN 采用的 C3D 网络结构如下图 3 所示。共具有 8 个卷积层、5 个池化层、两个全连接层，以及一个 softmax 输出层。所有 3D 卷积滤波器均为 3×3×3，步长为 1×1×1。为了在时间维度不过早的进行压缩，pool1 核大小为 1×2×2、步长 1×2×2，后面所有 3D 池化层均为 2×2×2，步长为 2×2×2。每个全连接层有 4 096 个输出单元。C3D 最终通过 softmax 层给出视频样本的分类。S-CNN 候选区域提取算法具体流程图见图 3。

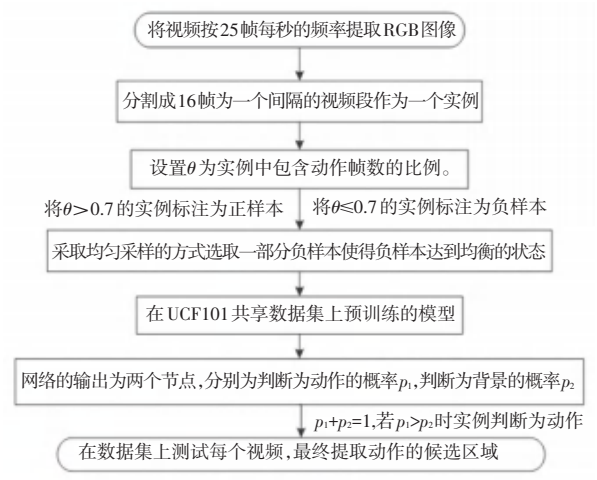


图 3　S-CNN 候选区域提取算法训练测试过程

Fig. 3　Training and testing process of S-CNN candidate region extraction algorithm

2.2.2　STFF-C3D CNN 候选区域提取算法

深度学习进行人体动作识别需要充分提取视频中的时间特征和空间特征，并合理的利用时空特征之间的相关性。为此，文献[11]提出一种采用稀疏采样方案的时空特征融合动作识别框架 STFF-C3D CNN（spatio-temporal feature fusion action recognition model，STFF-C3D CNN）。该模型不仅充分融合了视频中时空特征，并且运用稀疏采样方案[5]避免了冗余采样。主要分为 4 部分：稀疏采样生成 RGB 图和光流图、时空特征的提取、时空混合特征图的生成、C3D 卷积神经网络进行动作识别，如下图 4 所示。

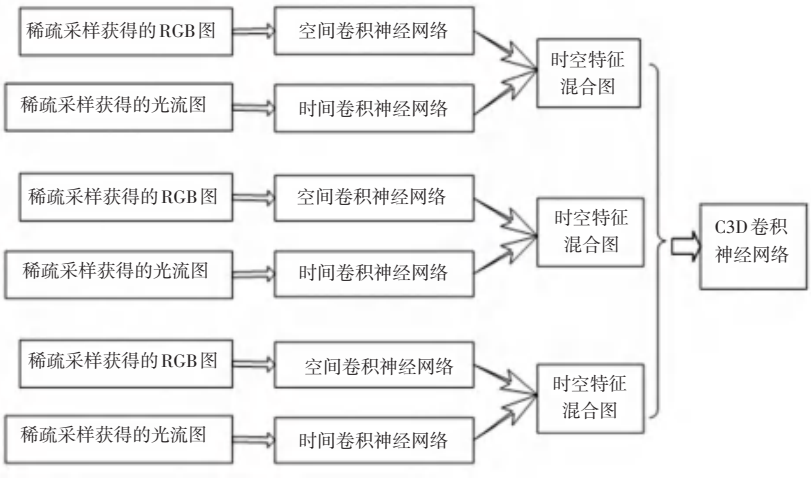


图 4　STFF-C3D CNN 框架

Fig. 4　STFF-C3D CNN frame

在预测阶段，建议片段（含有动作的片段）选择会影响到时序动作识别的效果和效率，若识别不准确，不仅影响识别结果的准确率，还会耗费时间去识别不准确的建议片段。本文提出的时空信息融合的

时序动作检测网络采用 STFF-C3D CNN[11] 进行建议片段的选择,可以充分融合候选片段的时间特征和空间特征来进行动作识别。

首先,采用稀疏采样方案对视频进行采样。一个输入视频被分为 K 段(segment),一个片段(snippet)从它对应的段中随机采样得到。对于每段 snippet,提取它包含空间信息的 RGB 图像和包含时间信息的 x 方向光流图和 y 方向光流图。接着,将 RGB 图送入空间卷积神经网络进行训练以提取中层空间特征,将光流图送入空间卷积神经网络进行训练以提取中层时间特征。然后,训练时空混合卷积神经网络提取时空融合特征,将时空中层特征图进行融合[14],提取混合卷积神经网络的中层混合特征,生成时空混合特征图。最后,将时空混合特征图作为 C3D 卷积神经网络的输入,在时间和空间维度分别进行卷积和池化,同时学习视频的运动信息和静态的图片信息,修改 STFF-C3D CNN 的输出为两类,即预测该片段是动作的概率以及是背景的概率。训练时将 IoU 大于 0.7 的作为正样本(动作),小于 0.3 的作为负样本(背景),对负样本进行采样使得正负样本比例均衡,采用 softmax loss 进行训练来判断动作类别。

2.3 3D 卷积反卷积神经网络

C3D 架构,由有 3 层全连接(FC)层的三维卷积神经网络组成,在诸如识别和定位等视频分析任务中取得了良好的结果。CDC 网络是基于 C3D 卷积神经网络构成的,C3D 的 conv1a 到 conv5b 是 CDC 网络的第一部分,对于 C3D 的其余层,CDC 保持 pool5 在长度和宽度上执行步长为 2 的最大池化,但保留了时间长度。按照常规设置[3,6,15],设置 CDC 网络输入的高度和宽度为 112×112,输入时间长度为 L 的视频段,pool5 的输出数据形式是(512, $L/8$, 4,4)。为了在帧级粒度上预测动作得分,需要在时间上进行上采样(从 $L/8$ 回到 L),在空间上进行下采样(从 4 × 4 到 1 × 1)。

CDC6 将卷积核设置为(4,4,4)、步长设置为(2,1,1)、填充设置为(1,0,0),因此 CDC6 可以将高度和宽度都减少到 1,同时将时间长度从 $L/8$ 增加到 $L/4$。CDC7 和 CDC8 都将卷积和设置为(4,1,1)、步长设置为(2,1,1)、填充设置为(1,0,0),因此 CDC7 和 CDC8 都进一步执行了步长为 2 的上采样,因此时间长度返回到 L。最后在 CDC8 的顶部添加帧级 SoftMax 层,以获得每个帧的置信度分数,每个通道代表一个类别。CDC 网络的最终输出形状为 (K + 1, L,1,1),其中 K + 1 代表 K 个动作类别加上背景类别。网络具体流程见下图 5。

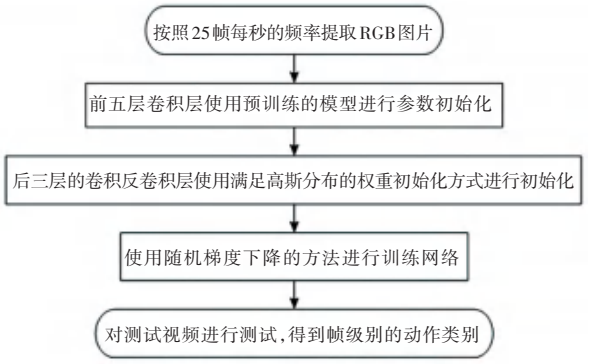


图 5 CDC 网络流程图

Fig. 5 CDC network flow chart

2.4 动作起始边界的调整

为了解决动作起始和终止坐标存在较大偏移的问题,本文提出了一个动作起始终止状态判断网络。改进文献[21]提出的网络,训练 Densnets 网络用作动作状态检测网络。Densnets 建立了不同层之间的连接关系,充分利用了特征,进一步减轻了梯度消失问题。训练测试过程如下:(1)首先,为了更大的间隔,将每个建议段的边界在两侧扩展了原始段长度的百分比 α。本文把所有实验的 α 设为 1/8。(2)对扩大后的时间区域提取光流图。(3)训练动作起始终止状态判断网络,得到判断模型。假设一个动作实例 S 的起始和终止坐标分别为 (t_1,t_2),则动作的时间长度 $L = t_2 - t_1$。训练过程如图 6 所示。(4)测试阶段,对测试所得的动作实例的 ($t_1 - L/8$, $t_1 + L/8$), ($t_2 - L/8$, $t_2 + L/8$) 中的帧进行测试,分别输出每一帧为开始帧,结束帧的概率,最后得到精修后的动作起始终止坐标。

3 实 验

THUMOS'14[16]:时间动作定位 THUMOS'14 数据集包含 20 类行动。本文用 2 755 个修剪的训练视频和 1 010 个未修剪的验证视频(3 007 个动作实例)来训练模型。测试使用的的 213 个不完全是背景视频的视频(3 358 个动作实例)。

评估指标:根据传统的度量标准[17],本文将每帧标记任务作为一个检索问题来处理,对于每个动作类,将测试集中的所有帧按其在该类中的置信度得分进行排序,并计算平均精度均值(AP),然后对所有的类进行平均得到平均 AP(mAP)。具有正确的预测类别且其与真实值的时间重叠度大于阈值时,预测是正确的,不允许对同一真实实例进行重复检测。

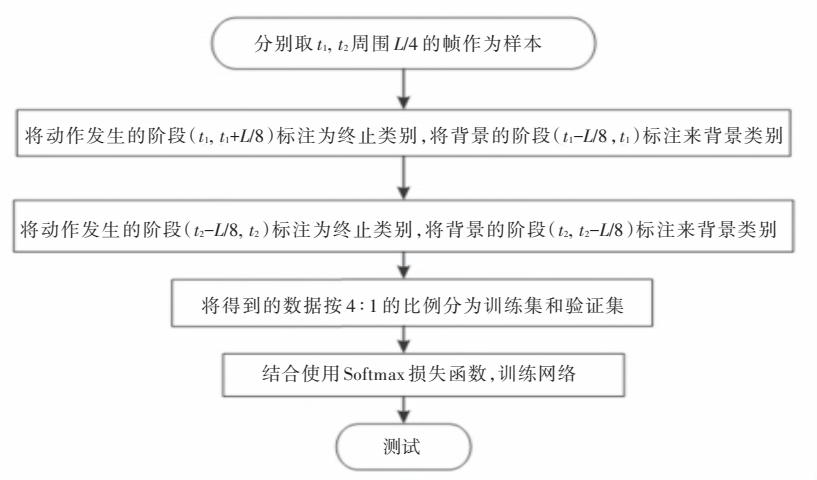


图 6　动作起始终止状态判断网络

Fig. 6　Action starting and ending state judgment network

每个测试结果都包含动作发生的时间区域和动作所属的类别.重叠度 IoU 计算为：

$$\mathrm{IoU}=\frac{R_p \cap R_{gt}}{R_p \cup R_{gt}}.$$

式中：R_p 表示预测的动作区域，R_{gt} 表示真实动作区域。如果重叠度 IoU 大于阈值，则表示预测是正确的。

理论上，因为卷积滤波器和 CDC 滤波器都在输入视频上滑动，所以它们可以应用于任意大小的输入。因此，CDC 网络可以在不同长度的视频上操作。但是由于 GPU 存储器的限制，实际上本文在视频上滑动 32 帧的时序无重叠窗口，并将每个窗口逐个送入 CDC 网络以及时获得时间上的密集预测。从时间边界的标签中，知道每个帧的标签，相同窗口中的帧可以有不同的标签。为防止包含太多背景帧进行训练，只保留至少有一帧属于动作的窗口。因此，在给定一组训练视频的情况下，能获得带有帧级标签的窗口训练集。

本文实验评估了在 THUMOS´14 数据集上重叠 IoU 阈值在 0.3~0.7 之间变化时的 mAP。如表 1 所示，CDC 获得的结果要好于所有其他先进方法的结果。与建议的 CDC 模型相比：用 FV 编码 iDTF 的系统[18]不能直接从原始视频中学习时空模式来进行预测。基于 RNN/LSTM 的方法（Yeung 等[19]，Yuan 等[20]）无法在时间依赖性之外明确捕获运动信息。S-CNN 可以有效地捕捉原始视频的时空模型，比其他 3 种方法的 mAP 得到了明显的提高，但是仅在段级粒度上进行时序动作定位，缺乏调整候选建议边界的能力。CDC 网络通过反卷积操作恢复时间维度上的长度，可以超出段级水平预测确定细粒度的置信度分数，在帧级粒度上进行时序动作定位的检测，因此可以精确定位时间边界。比 S-CNN 方法在各个阈值上的的 mAP 提高了 0.7% ~ 4.4%。

本文使用 STFF-CDC 网络，改进候选片段生成算法充分利用了视频的时空特征。实验结果表明，本文方法精度比 CDC 网络约提高了 2.2%。另外，本文采用 Densnets 网络作为动作状态检测网络，更有效地利用了特征，比 DSTIN+CEN[21] 方法相比也得到了提高。总之，本文模型在各个

阈值上均获得了其他方法更准确的时序动作定位效果。

部分检测结果对比展现在图 7，受益于候选区域选择算法的改进，和动作状态检测网络的贡献，本文的方法可以得到更为精确的动作区域。

表 1　本文动作识别算法和其他算法 mAP(%)对比

Tab. 1　The comparison of action recognition algorithm in this paper with others(map (%))

IoU 阈值	0.3	0.4	0.5	0.6	0.7
Richard and Gall[18]	30.0	23.2	15.2		
Yeung et al[19]	36.0	26.4	17.1		
Yuan et al[20]	33.6	26.1	18.8		
S-CNN[6]	36.3	28.7	19.0	10.3	5.3
CDC[10]	40.1	29.4	23.3	13.1	7.9
DSTIN+CEN[21]	44	34.8	25.8	16.6	8.3
STFF-CDC	46.2	36.9	27.1	17.8	9.7

图 7 部分检测结果对比

Fig. 7 Comparison of some test results

4 结束语

时序动作定位任务需要识别出一段长视频中的动作类别以及动作的起始时间。为了有效的提取候选区域,本文提出了一种基于时空特征融合的候选区域提取网络;为了提高动作起始点定位的精度,本文提出一种动作状态检测网络。在数据集 THUMOS'14 上进行实验,并与其他方法进行了对比。结果证明,本文提出的基于时空信息融合的时序动作定位模型可以有效进行时序动作定位,达到了较好的精度。

参考文献

[1] SIMONYAN K,ZISSERMAN A.Two-Stream convolutional networks for action recognition in videos [J]. Advances in Neural Informa-tion Processing Systems,2014,1(4):568-576.

[2] KARPATHY A, TODERICI G, SHETTY S, et al. Large Scale Video Classification with Convolutional Neural Networks [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2014,1725-1732.

[3] DU TRAN, BOURDEV L, ROB FERGUS, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proc of IEEE Inte - rnational Conference on Computer Vision. Piscataway,NJ:IEEE Press,2015:4489-4497.

[4] 胡珂杰, 蒋敏, 孔军. 基于混合关节特征的人体行为识别[J]. 传感器与微系统,2018,37(3):138-144.

[5] WANG LIMIN,XIONG YUANJUN,WANG ZHE,et al. Temporal segment networks: Towards good practices for deep action recognition [C]//Proc of European Conference on Computer Vision. Berlin: Springer,2016:20-36.

[6] SHOU Z, WANG D, CHANG S F. Temporal action localization in untrimmed videos via multi-stage cnns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1049-1058.

[7] ONEATA D, VERBEEK J, SCHMID C. The lear submission at thumos[J]. 2014.

[8] WANG R, TAO D. Uts at activitynet 2016[J]. AcitivityNet Large Scale Activity Recognition Challenge, 2016, 8: 2016.

[9] SINGH G, CUZZOLIN F. Untrimmed video classification for activity detection: submission to activitynet challenge[J]. arXiv preprint arXiv:1607.01979, 2016.

[10] SHOU Z, CHAN J, ZAREIAN A, et al. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5734-5743.

[11] 王倩,孙宪坤,范冬艳. 基于深度学习的时空特征融合人体动作识别[J]. 传感器与微系统.录用尚未发表.

[12] HUANG G, LIU Z, Van Der MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[13] 李崇国. 光流法在靶区运动图像配准中的应用研究[D]. 泸州医学院,2009.

[14] 杨天明,陈志,岳文静. 基于视频深度学习的时空双流人物动作识别模型[J]. 计算机应用,2018,38(3),895-899.

[15] TRAN D, BOURDEV L, FERGUS R, et al. Deep end2end voxel2voxel prediction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2016: 17-24.

[16] JIANG Y G, LIU J, ZAMIR A R G, et al . THUMOS challenge: Action recognition with a large number of classes [DS/OL]. (2014) http://crcv.ucf.edu/THUMOS14/.

[17] YEUNG S, RUSSAKOVSKY O, JIN N, et al. Every moment counts: Dense detailed labeling of actions in complex videos[J]. International Journal of Computer Vision, 2018, 126(2-4): 375-389.

[18] RICHARD A, GALL J. Temporal action detection using a statistical language model [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3131-3140.

[19] YEUNG S, RUSSAKOVSKY O, MORI G, et al. End-to-end learning of action detection from frame glimpses in videos[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2678-2687.

[20] YUAN J, NI B, YANG X, et al. Temporal action localization with pyramid of score distribution features[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3093-3102.

[21] 胡齐齐, 汪剑鸣, 金光浩. 基于时空信息的时序动作检测方法研究[J]. 微电子学与计算机, 2019, 36(2):94-98.