Journal of System Simulation

Volume 31 | Issue 11 Article 22

12-13-2019

A Temporal Action Detection Algorithm Based on Spatio-Temporal Feature Pyramid Network

Liu Wang
School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China;

Jinyu Sun
School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China;

Shiwei Ma
School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China;

A Temporal Action Detection Algorithm Based on Spatio-Temporal Feature Pyramid Network

Abstract

Abstract: In view of the discontinuity of motion timing detection in the frame-level prediction network structure, a novel algorithm based on spatio-temporal feature pyramid network (ST-FPN) is proposed. *In the frame-level action prediction, several 3D convolution-de-convolution (CDC) networks are used to sample spatial feature down to 1 dimension and sample temporal feature up to corresponding proposal level.* Then the prediction scores of different CDC networks are fused by non-maximum suppression (NMS). The softmax classifier is used to classify frame-level actions, and then temporal action detection is obtained. The experimental results on dataset THUMOS14 show that the proposed algorithm improves the accuracy of temporal action detection.

Keywords

temporal action detection, feature fusion, spatio-temporal feature pyramid network, 3D convolution-de-convolution, non-maximum suppression

Recommended Citation

Liu Wang, Sun Jinyu, Ma Shiwei. A Temporal Action Detection Algorithm Based on Spatio-Temporal Feature Pyramid Network[J]. Journal of System Simulation, 2019, 31(11): 2382-2387.

第 31 卷第 11 期 系统仿真学报© Vol. 31 No. 11
2019 年 11 月 Journal of System Simulation Nov., 2019

基于时空特征金字塔网络的动作时序检测方法

刘望，孙金玉，马世伟*
(上海大学 机电工程与自动化学院，上海 200444)

摘要：针对帧级别预测网络结构中的动作时序检测提案不连续问题，提出基于时空特征金字塔网络的动作时序检测算法。*在帧级别动作预测中，采用多个 3D 卷积反卷积网络，将空间特征维度降至 1 维，并将时间特征维度还原到相应的提案长度，得到不同时间尺度下的多个预测。*采用非极大值抑制的方式融合多个子网络的预测，并用分类器进行帧级别动作分类，进而得到时序提案。在共享数据集 THUMOS14 上的实验结果表明，该算法有效地提高了动作的时序区域定位精度。

关键词：动作时序检测；特征融合；时空特征金字塔；3D 卷积反卷积；非极大值抑制
中图分类号：TP391.4 文献标识码：A 文章编号：1004-731X (2019) 11-2382-06
DOI: 10.16182/j.issn1004731x.joss.19-FZ0369

A Temporal Action Detection Algorithm Based on Spatio-Temporal Feature Pyramid Network

*Liu Wang, Sun Jinyu, Ma Shiwei**
(School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China)

Abstract: In view of the discontinuity of motion timing detection in the frame-level prediction network structure, a novel algorithm based on spatio-temporal feature pyramid network (ST-FPN) is proposed. *In the frame-level action prediction, several 3D convolution-de-convolution (CDC) networks are used to sample spatial feature down to 1 dimension and sample temporal feature up to corresponding proposal level.* Then the prediction scores of different CDC networks are fused by non-maximum suppression (NMS). The softmax classifier is used to classify frame-level actions, and then temporal action detection is obtained. The experimental results on dataset THUMOS14 show that the proposed algorithm improves the accuracy of temporal action detection.

Keywords: temporal action detection; feature fusion; spatio-temporal feature pyramid network; 3D convolution-de-convolution; non-maximum suppression

引言

人体动作检测一直是计算机视觉领域的一个重要研究课题，受到了越来越多的关注。一般来说，动作检测的任务包括动作识别和动作定位 2 部分，前者的目的是回答视频中发生了什么动作，后者则致力于估计一个动作的开始和结束的时间。文献[1-8]展示了在人工剪辑的短视频中识别动作已经取得的进展，但在长时间的未修剪视频中检测动作即动作时序检测，仍然是一个具有挑战的难题。

动作时序检测任务不仅需要判断视频中动作的类别，还需要推断出动作发生的时间范围。文献[9-14]提出了各种各样的模型以处理动作时序检测中的问题，其中一个难点就是动作存在不同尺度。

收稿日期：2019-05-21 修回日期：2019-07-23；
基金项目：新疆兵团重大项目子项目(2018AA008-04)；
作者简介：刘望(1995-)，男，福建福州，硕士生，研究方向为视频检索；马世伟(通讯作者 1965-)，男，甘肃嘉峪关，博士，教授，研究方向为信号处理、图像处理和模式识别等。

在未剪辑的视频中，动作可能持续几帧也可能长达数秒。文献[9]提出了一个 3 阶段的 3D 卷积网络(Segment-CNN，S-CNN)，通过提案网络、分类网络和定位网络进行动作时序检测，但其中的 C3D[6]架构使输出特征的时序长度减少到 1 帧，缺乏对持续时间较短的动作的捕获能力。文献[10]在 S-CNN 的基础上提出了时空卷积反卷积结构(convolution-de-convolution，CDC)，将动作时序检测的细粒度提升到帧级别，但丢失了高层次特征信息，导致动作判别的连续性降低。

本文基于 CDC 结构提出一种时空特征金字塔网络(spatio-temporal feature pyramid network，ST-FPN)，通过多个反卷积网络，对长度为 L 帧的视频片段独立做出 $L/4$ 段、$L/2$ 段、L 段的预测。并将帧级别预测与段级别预测通过上采样和非极大值抑制进行融合，得到最终预测结果。实验表明该方法解决了帧级别预测存在的提案不连续问题，在共享数据集 THUMOS14[11]上，检测精度得到了提高。

1 相关算法

1.1 3D 卷积

卷积可以直接对输入数据进行特征提取，实现端到端学习[15]。三维卷积是在每 3 或 4 个连续帧上应用时空滤波器。利用该机制，卷积层不仅可以提取单帧图像中的空间信息，还可以学习帧间的时间信息，完成时空特征提取。因此，在 3D 卷积中，第 j 层上点(x, y, z)的特征由第 i 层特征和时空滤波器卷积得到，如式(1)所示：

$$v_{ij}^{xyz} = f\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}\omega_{ijm}^{pqr}v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \quad (1)$$

式中：R_i 为 3D 卷积核的时间维度；ω_{ijm}^{pqr} 为第 m 个特征图上的第(p, q, r)个权重矩阵。文献[6]提出了一种基于 3D 卷积和 3D 池化的视频特征识别结构(convolution 3D，C3D)，在视频任务中被广泛使用，C3D 模型结构如图 1 所示。

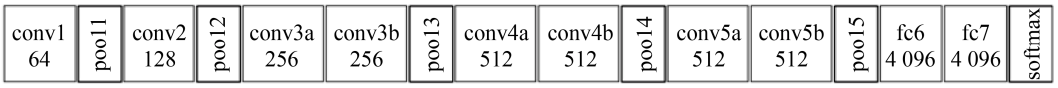


图 1 C3D 模型结构
Fig. 1 C3D model architecture

1.2 3D 反卷积

反卷积结构最早应用于图像任务中代替上采样层填充图像的内容，在图像分割、图像生成等任务中取得了引人注目的效果[16]。文献[10]中的 3D 卷积反卷积结构(CDC)有别于 2D 卷积反卷积结构。它仅在时间上进行上采样，而在空间上进行降采样，从而实现帧级别的输出。CDC 结构如图 2 所示。

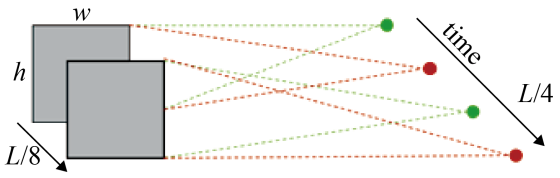


图 2 3D 卷积反卷积滤波器示意图
Fig. 2 Illustration of 3D convolution-de-convolution filter

定义一个 CDC 结构中滤波器 F 的大小为(k_l, k_h, k_w)，输入的 X 的空间大小为(k_h, k_w)，输出 Y 包括了时间维度上的 k_l 个连续输出。那么对于给定的空间坐标 $a \in \{1,2,\ldots,k_h\}$，$b \in \{1,2,\ldots,k_w\}$ 及时间坐标 $c \in \{1,2,\ldots,k_l\}$，CDC 的前向传播可以通过式(2)求出。

$$Y[c] = \sum_{a=1}^{k_h}\sum_{b=1}^{k_w}F[c,a,b]\cdot X[a,b] \quad (2)$$

同样地，CDC 的反向传播同样遵循链式法则，通过 Y 将梯度传播给 X，如式(3)所示：

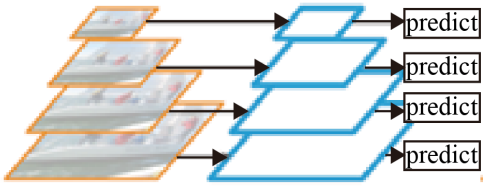
$$X[a,b] = \sum_{c=1}^{k_l}F[c,a,b]\cdot Y[c] \quad (3)$$

CDC 滤波器 F 是一系列全连接卷积核的组合，其中的每一个卷积核都具有和输入 X 相同的内核

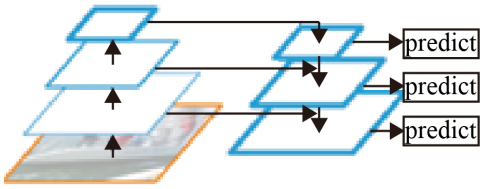
大小(k_h, k_w)，并共享输入域 X，使 CDC 具有空间维度降采样的功能。同时，在时间维度上的 k_l 个连续输出，使 CDC 结构具备时间维度上采样的功能。

1.3 特征金字塔

图像金字塔是多尺度目标检测系统的基本组成部分，但图像金字塔需要在每个图像尺度上对特征独立计算，速度较慢[17]，如图 3(a)所示。文献[17]提出了一种具有横向连接的自顶向下的结构，用于构建各种尺度的高级语义特征图，称为特征金字塔网络(feature pyramid network，FPN)，如图 3(b)所示。顶层的特征可以有效地预测大尺度的信息，底层的特征则对小尺度目标更为敏感。



(a) 图像金字塔



(b) 特征金字塔

图 3 特征金字塔网络示意图
Fig. 3 Illustration of feature pyramid network

2 本文方法

为有效地结合全局预测和局部预测的信息，提出一种时空特征金字塔网络(ST-FPN)。如图 4 所示，网络主要由 3 部分组成：(1) 3D 卷积特征提取网络；(2) 时空特征金字塔网络；(3) 动作提案。

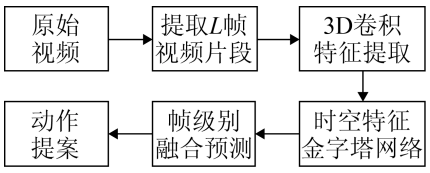


图 4 动作时序检测框架
Fig. 4 Framework for temporal action detection

2.1 视频动作特征提取

本文基于 C3D 网络提取视频序列中的动作特征。C3D 网络可以同时学习到动作的外观和运动特征。首先从未剪辑的视频 v 中提取出一段时间长度为 L 的视频序列，记为 V。V 的大小为 $c*l*h*w$，其中 c 为每帧图片的通道数，本文使用 RGB 图像，因此 c 为 3；l 为输入网络的视频序列的帧数；w 和 h 表示视频帧的宽度和高度。其次，将 V 输入 C3D 网络得到第 5 层的输出特征。由于 C3D 网络中的 5 次池化，V 的时间长度缩短为 L/8。其中第一次池化和最后一次池化的时间步长均设置为 1，以便网络保留更多的细节信息。3D 卷积特征提取网络的具体网络参数如表 1 所示。

表 1 3D 卷积网络参数
Tab. 1 Parameters of 3D convolution network

层	核大小	输出大小
input	原始输入 3×L×112×112	
conv1	3×3×3	64×L×112×112
pool1	3×2×2 stride (1, 2, 2)	64×L×56×56
conv2	3×3×3	128×L×56×56
pool2	3×2×2 stride (2, 2, 2)	128×L/2×28×28
conv3_x	$\begin{bmatrix}3\times3\times3\\3\times3\times3\end{bmatrix}$	256×L/2×28×28
pool3	3×2×2 stride (2, 2, 2)	256×L/4×14×14
conv4_x	$\begin{bmatrix}3\times3\times3\\3\times3\times3\end{bmatrix}$	512×L/4×14×14
pool4	3×2×2 stride (2, 2, 2)	512×L/8×7×7
conv5_x	$\begin{bmatrix}3\times3\times3\\3\times3\times3\end{bmatrix}$	512×L/8×7×7
pool5	3×2×2 stride (1, 2, 2)	512×L/8×4×4

2.2 时空特征金字塔网络

针对帧预测的不连续性问题，提出了时空特征金字塔网络(ST-FPN)。ST-FPN 在一次卷积特征提取操作中对不同长度的动作片段进行预测，并在时间维度上融合，满足了不同尺度动作检测的要求。ST-FPN 由 3 个子网络 A，B，C 组成，每个子网络都独立生成不同尺度的预测得分。之后，通过非极大值抑制剔除较小的得分，进而预测动作概率。

如图 5 所示，视频序列 V 首先通过 3D 卷积得到时间维度为 $L/8$ 的特征，再通过 3 层不同的 3D 卷积反卷积结构得到不同尺度的特征。

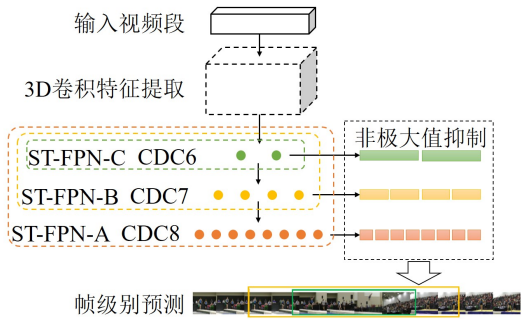


图 5 时空特征金字塔网络示意图

Fig. 5 Illustration of spatio-temporal feature pyramid network

其中，ST-FPN-A 网络通过 3 次 CDC 操作得到，每次 CDC 操作会将特征时间维度扩大两倍，使时间维度为 $L/8$ 的特征，还原到维度为 L 的特征，即原 CDC 网络；ST-FPN-B 网络通过 2 次 CDC 操作得到，将时间维度还原为 $L/2$；ST-FPN-C 网络仅执行 1 次 CDC 操作，得到的时间维度为 $L/4$。

对于 ST-FPN-C、ST-FPN-B 得到的高、中层次语义特征通过上采样将其映射回 L 维。并在时间维度上将 3 个子网输出融合，即通过非极大值抑制剔除较小的预测得分得到融合结果。通过 softmax 分类器进行分类。3 个子网详细参数如表 2 所示。

表 2 时空特征金字塔网络参数

Tab. 2 Parameters of spatio-temporal feature pyramid network

输入	3D 卷积特征 512×L/8×4×4					
子网络	ST-FPN-A		ST-FPN-B		ST-FPN-C	
层	核大小	输出大小	核大小	输出大小	核大小	输出大小
CDC6	4×4×4 (2 copies)	4 096×L/4×1×1	4×4×4 (2 copies)	4 096×L/4×1×1	4×4×4 (2 copies)	(K+1)×L/4×1×1
CDC7	4×1×1 (2 copies)	4 096×L/2×1×1	4×1×1 (2 copies)	4 096×L/2×1×1	-	-
CDC8	4×1×1 (2 copies)	(K+1)×L×1×1	-	-	-	-
上采样	1	(K+1)×L×1×1	2	(K+1)×L×1×1	4	(K+1)×L×1×1
NMS	(K+1)×L×1×1					
softmax	(K+1)×L×1×1					

其中，CDC6 层卷积核大小为 $(k_{cin}, k_w, k_h) = (4, 4, 4)$，高度和宽度减小到 1；2 copies 表示卷积核数量 $k_l = 2$，使输出时间维度从 $L/8$ 扩大到 $L/4$；对于 A、B 网络，输出通道维度都为 $k_{cout} = 4\,096$，因此共享 CDC6 层，以减少网络参数；对于 C 网络，输出通道维度 $k_{cout} = K + 1$，即直接输出 $K+1$ 个类别的预测(K 个动作类，1 个背景类)。CDC7 和 CDC8 层参数与 CDC6 层类似，此处不再赘述。此外，CDC 结构中所有卷积核步长设置为(2, 1, 1)填充为(1, 0, 0)。

在上采样层中，A，B，C 网络分别执行 1，2，4 倍的上采样使输出时间维度统一为 L。NMS 层将不同子网络上采样后的时间维度对齐，按帧和类别进行非极大值抑制(NMS)，如式(4)所示。

$$p_{k,l} = \max\{p_{A,k,l}, p_{B,k,l}, p_{C,k,l}\} \tag{4}$$

式中：$p_{A,k,l}$，$p_{B,k,l}$，$p_{C,k,l}$ 分别为经过上采样后 A，B，C 网络的第 k 类第 l 帧的预测得分，$p_{k,l}$ 为非极大值抑制融合得分的结果。

2.3 动作提案

对于时空特征金字塔预测的结果，本文使用连续预测生成提案的方法，即预测出相同类别的连续帧生成一个动作提案段，提案段的置信概率通过计算该段所有帧的置信概率均值得到。为简化计算，本文只记录动作帧 TOP1 置信概率的提案，同时删除背景提案和长度小于 0.1 s 的提案。

3 实验

3.1 数据集

实验选取共享数据集 THUMOS14 对本文方法

DOI: 10.16182/j.issn1004731x.joss.19-FZ0369

第 31 卷第 11 期 系统仿真学报 Vol. 31 No. 11
2019 年 11 月 Journal of System Simulation Nov., 2019

进行评估。THUMOS14 验证集和测试集分别包含 1 010 和 1 574 个未剪辑的视频。在动作时序检测任务中，只有 20 个动作被标记，包括 200 个验证视频(3 007 个动作片段)，213 个测试视频(3 358 个动作片段)。每个视频都包含一个或多个特定动作。

3.2 实验设置

实验基于一台包含 TITAN Xp 显卡的工作站，在 3D-CAFFE 框架下运行。实验探索了 3 个子网与部分子网组合的结果，使用 C3D[6]预训练模型，采用小批量随机梯度下降的方法进行训练，训练时学习率设置为 0.000 01，批量大小设置为 8。此外，采样率为 25 fps，训练阶段输入为 32 帧，测试阶段输入分别为 32 帧、64 帧和 128 帧。

3.3 实验结果

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，在不同交并比(intersection over union，IoU)阈值下计算检测精度，即平均精度均值(mean average precision，mAP)。IoU 计算方式为：

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

式中：R_p 为预测动作区间；R_{gt} 为真实动作区间。如果 IoU 大于阈值，则表示预测正确。实验计算了不同阈值下各网络的表现。首先针对 32 帧的输入进行测试，结果如表 3 所示。其中 ST-FPN-A(即 CDC 网络)为实验的基准线，ST-FPN-A+B 为 A、B 为子网络进行融合的结果，ST-FPN-A+B+C 为所有子网络融合的结果。实验表明，A 网络对细节信息有较好的捕捉，IoU 阈值为 0.1 的预测精度较高；B 网络对中层特征的预测精度较好；C 网络结果不佳，分析认为是由于3D卷积得到的特征维度仅512维(A、B 网先进行一次卷积反卷积操作，特征维度提升至 4 096 维)，直接预测无法有效区分样本特征。因此，在融合网络的比较中，A、B、C 网络的融合结果受 C 网络的影响，精度劣于仅基于 A、B 网络融合的结果。同时，融合网络的精度仅在 IoU 阈值为 0.1 处较好是因为 32 帧输入区间较短，子网并不能很好地判断出全局信息。

表 3 不同网络检测精度比较(32 帧输入)
Tab. 3 Comparation of detection accuracy among different networks (32-frame input)

IoU threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7
ST-FPN-A	**39.1**	26.1	16.4	8.8	4.8	2.4	1.1
ST-FPN-B	38.2	**28.0**	**19.9**	**13.5**	**8.0**	**4.5**	**2.1**
ST-FPN-C	15.9	8.3	4.9	2.7	1.5	0.8	0.3
ST-FPN-A+B	**39.8**	**27.6**	**18.9**	**12.2**	**6.7**	**3.9**	**1.4**
ST-FPN-A+B+C	21.4	12.6	7.8	4.5	2.3	1.0	0.4

基于上述原因，继续对 64 帧和 128 帧的输入进行实验，且实验仅记录 A、B 子网络及 A、B 联合网络的结果。如表 4 所示，每一个 IoU 阈值下都有 2 个分别对应 64 帧和 128 帧输入的检测精度，最优的结果加粗标注。可以看出，随着输入区间的增大，各网络的检测精度都得到提升。并且，联合网络的检测精度基本上优于单独子网的预测，证明了时空特征金字塔通过融合不同时间尺度的动作预测可以解决帧级别预测不连续的问题。

此外，本文选取了一些现有方法来进行对比。如表 5 所示，IoU 阈值小于 0.4 时的检测精度表现更好，即低层次动作的时间区域得到更为精确的捕捉，证明了时空特征金字塔对不同尺度分别预测再融合的方法可以优化帧级别时序检测的不连续性问题。此外，相比于 S-CNN[9]方法需要同时对 16-512 个帧长度的输入域进行卷积来得到不同尺度的预测，本文的方法仅进行一次卷积操作即可完成不同尺度的预测，效率更高，也更具有竞争力。

表 4 不同网络检测精度比较(64 帧，128 帧输入)
Tab. 4 Comparation of detection accuracy among different networks (64/128-frame input)

IoU threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7
ST-FPN-A	43.3/45.1	32.2/35.7	20.5/23.6	11.3/13.8	5.9/6.8	3.0/3.6	1.3/1.6
ST-FPN-B	43.3/44.3	33.1/35.3	22.8/26.3	15.7/18.5	**10.3**/12.1	**5.9**/6.9	2.8/**3.7**
ST-FPN-A+B	**47.1**/**50.4**	**36.4**/**41.3**	**25.6**/**30.3**	**16.6**/**20.7**	10.0/**13.2**	5.7/**7.5**	**2.8**/3.5

表 5 本文方法与其他方法的检测精度比较
Tab. 5 Comparation of detection accuracy between the proposed method and other methods.

IoU threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Karaman 等	4.6	3.4	2.1	1.4	0.9	-	-
Wang 等[11]	18.2	17.0	14.6	12.1	8.5	4.7	1.5
Oneata 等[12]	36.6	33.6	27.0	20.8	14.4	8.5	3.2
Richard 和 Gall[13]	39.7	35.7	30.0	23.2	15.2	-	-
Yeung 等[14]	48.9	44.0	36.0	26.4	17.1	-	-
Shou 等[9] (S-CNN)	-	-	36.3	28.7	19.0	10.3	5.3
ST-FPN-A+B (L=128)	50.4	41.3	30.3	20.7	13.2	7.5	3.5

4 结论

本文提出一种新颖的网络架构，通过空间特征金字塔的方式，融合了不同尺度下的动作预测，解决了帧级别预测的不连续性问题，更准确地预测了动作发生的起止时间及类别。在共享数据集 THUMOS14 中的实验结果表明，本文的方法在精度或效率上得到了显著的提升。在未来的工作中，将改进网络结构，提高高层语义特征的检测精度，从而提高整体动作时序检测精度。

参考文献:

[1] Herath S, Harandi M, Porikli F. Going Deeper into Action Recognition: A survey[J]. Image and Vision Computing (S0262-8856), 2017, 60: 4-21.

[2] Wang H, Schmid C. Action Recognition with Improved Trajectories[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3551-3558.

[3] 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. 2014: 1725-1732.

[4] Simonyan K, Zisserman A. Two-stream Convolutional Networks for Action Recognition in Videos[C]. Advances in Neural Information Processing Systems. 2014: 568-576.

[5] Ji S, Xu W, Yang M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (S0162-8828), 2012, 35(1): 221-231.

[6] 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.

[7] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised Learning of Video Representations using LSTMs[C]. International Conference on Machine Learning. 2015: 843-852.

[8] Carreira J, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6299-6308.

[9] Shou Z, Wang D, Chang S F. Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1049-1058.

[10] Shou Z, Chan J, Zareian A, et al. CDC: Convolutional-de-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5734-5743.

[11] Wang L, Qiao Y, Tang X. Action Recognition and Detection by Combining Motion and Appearance Features[J]. THUMOS14 Action Recognition Challenge, 2014, 1(2): 2.

[12] Oneata D, Verbeek J, Schmid C. The Lear Submission at Thumos 2014[J]. 2013.

[13] Richard A, Gall J. Temporal Action Detection Using a Statistical Language Model[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3131-3140.

[14] Yeung S, Russakovsky O, Mori G, et al. End-to-end Learning of Action Detection from Frame Glimpses in Videos[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2678-2687.

[15] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[C]. European Conference on Computer Vision. Springer, Cham. 2014: 818-833.

[16] Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional Networks[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2010, 10: 7.

[17] Lin T Y, Dollár P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

https://dc-china-simulation.researchcommons.org/journal/vol31/iss11/22
DOI: 10.16182/j.issn1004731x.joss.19-FZ0369