DOI：10.13245/j.hust.181207

一种融合全局时空特征的 CNNs 动作识别方法

王 珂 武 军 周天相 李瑞峰

(哈尔滨工业大学机器人技术与系统国家重点实验室，黑龙江 哈尔滨 150001)

摘要 针对基于卷积神经网络(CNNs)的人体动作识别方法通常采用空域或时域局部特征的不足，提出一种融合人体动作全局时域和空间特征的双通道 CNNs 动作识别模型．空间通道对动作图像进行深度学习，采用多帧融合的方式提升准确率，全局时域通道对能量运动历史图(EMHI)进行深度学习，最后融合两个通道信息识别人体动作．利用现有的大型数据集进行预训练，以解决学习过程中训练样本不足问题．在 UCF101 数据集和该项目小样本数据集上进行实验，结果证明了该方法的有效性．

关键词 动作识别；卷积神经网络；能量运动历史图；全局时域特征；数据集

中图分类号 TP391.41 **文献标志码** A **文章编号** 1671-4512(2018)12-0036-06

An action recognition method based on global spatial-temporal feature convolutional neural networks

Wang Ke Wu Jun Zhou Tianxiang Li Ruifeng

(State Key Laboratory of Robotics and System，Harbin Institute of Technology，Harbin 150001，China)

Abstract The existing human motion recognition methods based on convolution neural network (CNNs) usually use spatial or temporal local features．In this paper，a two-stream CNNs action recognition model was proposed，which integrated the global temporal and spatial features of human action．Motion images were deeply studied in spatial channels，the multi frame fusion way was used to raise the accuracy rate，and deep learning on the energy motion history image (EMHI) was performed in the global temporal stream．Finally，the two streams were combined to identify the human motion．In order to solve the problem of insufficient training samples in the learning process，the existing large data sets was used for pre-training．Experiments were carried out on the UCF101 dataset and the small sample dataset of the project．The results demonstrate the effectiveness of the method.

Key words action recognition；convolutional neural networks；energy motion history image；global temporal feature；dataset

由于在人机交互、智能交通系统、视频监控等多个领域的巨大需求，人体的动作识别越来越受到计算机视觉领域的重视．为了使计算机能识别来自不同场景的动作，其核心是利用判别特征来表征动作，然后对其进行分类．与静态图像识别不同，除了动作的空间特征外还有更为重要的动作的时间特征[1]，如何有效提取动作的空间特征和动作的时间特征是人体动作识别要解决的两个主要问题．

随着卷积神经网络(CNNs)在图像分类领域取得巨大成功，人们尝试从原始图像通过多层的卷积层和池化层自动学习动作特征．与图像分类相比动作具有动作的时间特征，用于动作识别的 CNNs 通常会比较复杂．大多数基于 CNNs 的动作识别方法[2-3]按照两个步骤来实现：首先利用静态图像建立空间 CNNs，然后在时间上将它们融合，这就导致动作之间的时间关系丢失．文献[1]提出一种 Two-stream CNNs 架构，通过当前静态图像来学习动作的动作的空间特征，通过帧间光流来学习动作的动

收稿日期 2018-06-15．

作者简介 王 珂(1979-)，男，讲师，E-mail：wangke@hit.edu.cn．

基金项目 国家自然科学基金资助项目(61673136)；教育部-中国移动科研基金资助项目(MCM20170208)．

作的时间特征，表明动作的时间特征具有更多的判别力．文献[4]在 Two-stream CNNs 的基础上，利用 CNN 网络进行了时间与空间上的融合，并将基础网络都换成了 VGG16．文献[5]提出一种时间分割网络(temporal segment networks，TSN)，将稀疏时间采样策略和基于视频的监督相结合，使用整个视频支持有效的学习．文献[6]设计了 3D-CNNs 架构，提出通过 3D 卷积核去提取视频数据的时间和空间特征，这些 3D 特征提取器在空间和时间维度上操作，因此可以捕捉视频流的运动信息，但是精度较低．文献[7]提出 C3D，采用 3D 卷积和 3D Pooling 构建网络．此外还有一些基于人体骨骼点来识别人体动作的方法，如文献[8]提出时空图卷积网络(spatial-temporal graph convolutional networks，ST-GCN)，设计不同的划分规则将每个节点附近的节点划分为不同的子集，对时空骨架序列进行图卷积操作．由于缺乏足够的数据集进行训练[9]，导致基于 CNNs 学习特征的动作识别方法相较于基于手工提取特征的动作识别方法准确率更低，仍有很大的提升空间．

本研究建立了包含空间通道和全局时域通道的双通道卷积神经网络结构，对人体动作进行识别．其中空间通道 CNNs 对动作图像进行深度学习，全局时域通道对本文提出的能量运动历史图(energy motion history image，EMHI)进行深度学习．考虑到对于某些动作缺乏数据集的问题，本研究采用现有的大型动作数据集来训练这些数据集较少的动作．

1　双通道卷积神经网络

基于分解假设[1]，可以将动作分解为空间和时域通道，但用帧间光流堆叠的方式只能表示动作的局部时域特征，忽略了动作的全局时空联系．本研究设计了一种包含空间和全局时域两个通道的卷积神经网络算法，对人体动作进行表征和识别，其流程图如图 1 所示．空间通道 CNNs 对动作图像进行深度学习，全局时域通道对能量运动历史图进行深度学习，最后通过两个通道进行融合．在空间通道

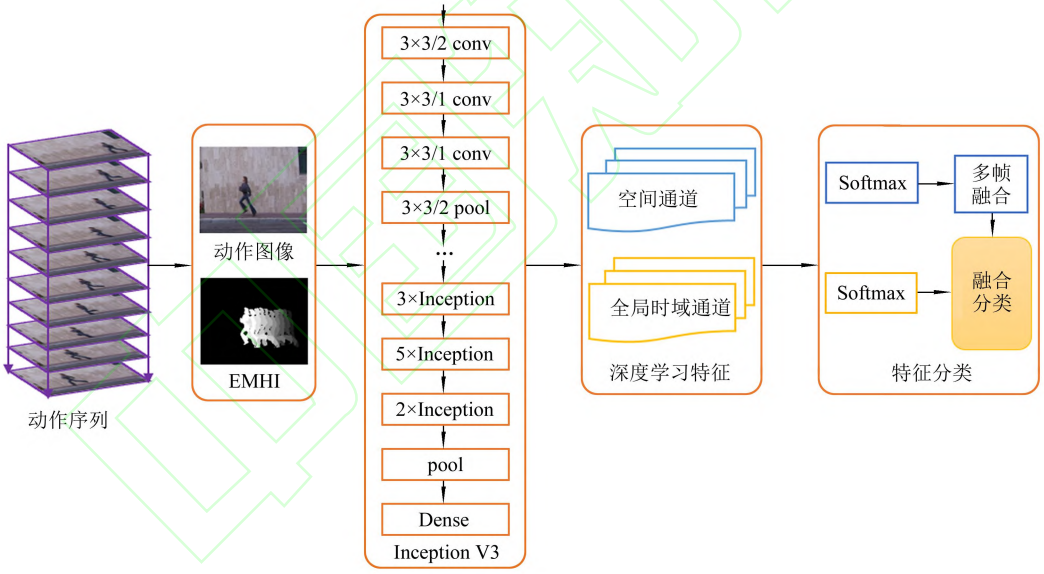

图 1　基于双通道卷积神经网络的动作识别流程图

采用多帧融合的方式进行动作识别，对当前帧和之前固定帧数的识别结果平均融合，提升识别准确率．

1.1　空间通道卷积网络

空间通道卷积网络的输入是单帧图像，这样的分类网络其实有很多，例如 GoogLeNet[10]，VGG[11]，ResNet[12]等，可以在 ImageNet[13]上预训练，再进行参数迁移．

CNNs 往往会随着深度加深而难以训练，VGG16 的 Keras model 为 588 MB，参数过多而且训练速度较慢．以往大多是对神经网络的结构在“宽度”和“深度”两个方面进行调整，而 ResNet50 则通过叠加大量浅层网络的方式来改善随着网络深度加深难以训练的情况．InceptionV3[14]通过将一个较大的二维卷积拆成两个较小的一维卷积，节约了大量参数，加快了运算速度而且减轻了过拟合，从而达到提高性能而又不大幅增加计算量的目的．而 ResNet50 和 InceptionV3 的 Keras model 仅有 100 MB 左右．经实验对比后发现 InceptionV3 在本研究的数据集上表现更好，所以选用 InceptionV3 为基础网络结构．

因为动作是一个三维的时空信号，若空间通道只以当前帧的输出作为判别依据，则可能出现较大

误差，所以本研究在空间通道采用多帧融合的方式进行动作识别，对当前帧和之前固定帧数的识别结果加权平均．如图 2 所示，将当前帧与前 2 帧的输出融合，虽然当前帧的识别出现错误，但通过前 2 帧的矫正最终输出了正确的结果，提高了识别的准确率．

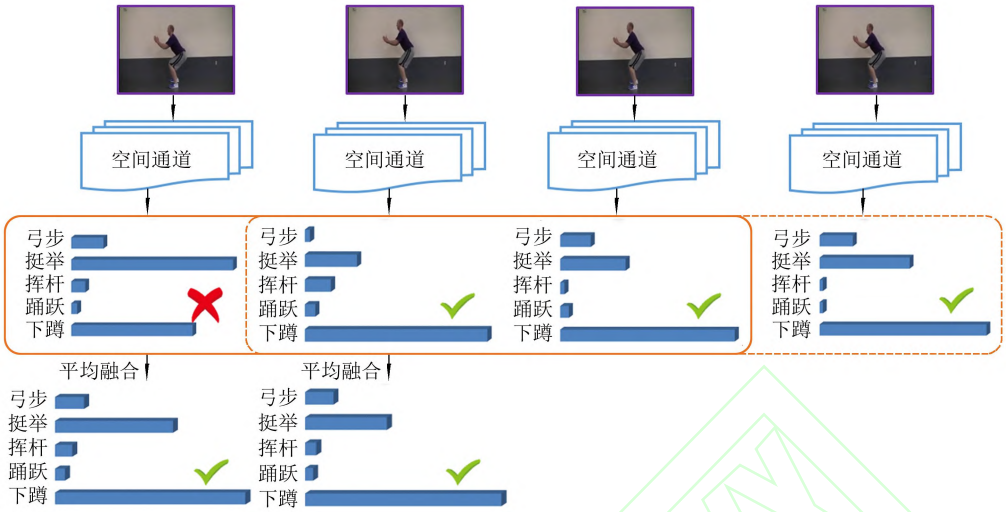

图 2　空间通道多帧融合示意图

1.2　全局时域通道卷积网络

文献[15]在运动能量图(motion energy image，MEI)的基础上提出运动历史图(motion history image，MHI)[16]来表征动作．MHI 是一种基于视觉的模板，通过计算一段时间内同一位置的像素变化，将人体动作用图像灰度值的形式表现出来．MHI 是一幅灰度图像，每个位置的灰度值代表在视频序列中该位置最近的运动情况，越接近当前帧的动作该位置的灰度值越大．

设 H 为 MHI 的灰度值，按照更新函数得到

$$H_\tau(x,\ y,\ t)=\begin{cases}\tau & (\psi(x,\ y,\ t)=1);\\ \max(0,\ H_\tau(x,\ y,\ t-1)-\delta) & (\text{其他}),\end{cases}$$

式中：$H_\tau(x,\ y,\ t)$ 为第 t 帧静态图像对应 MHI 中的坐标为$(x,\ y)$的像素点的灰度值；τ 为持续时间，决定了运动的时间范围；δ 为衰退参数；$\psi(x,\ y,\ t)$ 为更新函数，用于判断各个像素点在当前帧是否为前景，若为前景则等于 1．$\psi(x,\ y,\ t)$ 可由帧间差分法得到，

$$\psi(x,\ y,\ t)=\begin{cases}1 & (D(x,\ y,\ t)>\xi),\\ 0 & (\text{其他});\end{cases}$$

$$D(x,\ y,\ t)=\left|I(x,\ y,\ t)-I(x,\ y,\ t-\Delta)\right|,$$

式中：$I(x,\ y,\ t)$ 为第 t 帧图像位于 $(x,\ y)$ 坐标的像素点的灰度值；Δ 为帧间距离；ξ 为用来判别前景和背景的阈值．

考虑到很多动作是跨越很多帧的，如图 3 所示，在做蹲起运动的过程中，可能会保持下蹲姿势超过 10 帧以上，若用每一帧来更新 MHI 则会损失动作的全局时域特性．

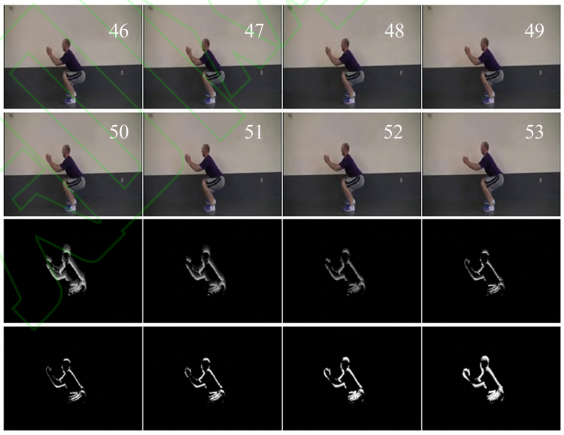

图 3　蹲起动作的运动历史图(MHI)

要对每个动作得到能够表示出在时域上具有全局性的 MHI，有必要设计一种自适应的方法．本研究提出一种能量运动历史图，其原理是通过判断当前帧与前一有效帧之间的运动能量来判断当前帧是否为有效帧，然后更新 EMHI．设 E_t 为第 t 帧相对于前一个有效帧 t_e 的运动能量，定义如下

$$E_t=\frac{1}{C}\sum_{(x,y)\in P}d(x,\ y,\ t);$$

$$d(x,\ y,\ t)=\sqrt{d^{\mathrm{u}}(x,\ y,\ t)^2+d^{\mathrm{v}}(x,\ y,\ t)^2},$$

式中：$d^{\mathrm{u}}(x,\ y,\ t)$ 和 $d^{\mathrm{v}}(x,\ y,\ t)$ 为像素点水平和竖直方向的光流；C 为有位移的像素点的个数；P 为等间距抽取的像素点集合用以计算稀疏光流；$d(x,\ y,\ t)$ 为像素点的位移．实质上是通过像素点的位移大小来判断是否为有效帧，但只是通过求图像内所有像素点的位移之和是不可行的．由于视角不同，运动的人物在图像中的比例是不同的，距离镜头近的人物做一个微小动作就可能得到很大的运

动能量，因此通过除以有效像素的个数来消除视角的影响．如果每帧都计算稠密光流，那么很难满足实时性要求，所以本研究通过等间距的稀疏光流来计算两帧之间的运动能量．

普通的 MHI 难以得到具有全局性的时域动作特征，而本文提出的 EMHI 会保留多帧以前的运动状态，从而得到更好的全局时域动作特征．如图 4 所示，上方的两行图片为有效帧，若当前帧为有效帧则更新 EMHI，反之不更新．人体在 42 到 53 帧之间保持下蹲姿势，身体只有微小的动作，EMHI 仍然可以得到较好的全局时域特征．

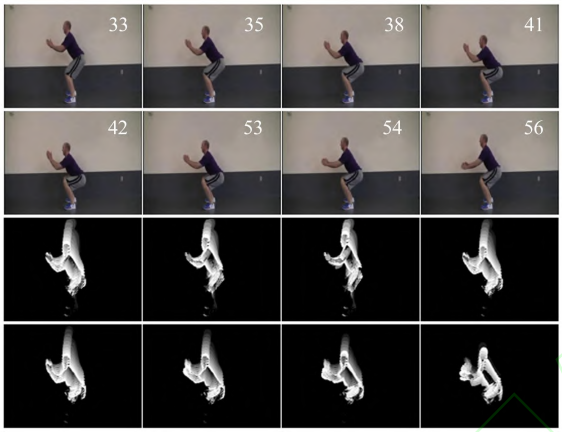

图 4　蹲起动作的能量运动历史图(EMHI)

2　小样本训练方法

基于视觉的人体动作识别方法须要以公共的动作数据集为前提和基础，当前比较常用的数据集有 UCF101 动作数据库和 HMDB 动作数据库等．虽然这些数据库中包含了较多的人体动作类别，但在实际应用中对于某些特定动作的样本很少．例如在家庭环境下判断老人是否摔倒、婴儿是否摔倒等动作的数据样本很少，不足以用来训练卷积神经网络．

文献[17]对传统 CNNs 模型进行了可视化的研究，发现 CNNs 前面的卷积层提取到的是一些关于边缘、条纹以及颜色的信息，之后的卷积层则会提取到一些结构信息．考虑到在人体动作识别中 CNNs 的卷积层可能会提取到一些运动的基本特征，所以根据现有的大型数据集来预训练模型，最后迁移至这些小样本的动作用于识别．

2.1　小样本数据集

本研究针对摔倒、走路、坐下、站立 4 个动作进行了数据集的采取，弥补了项目在数据集上的空白．为了使模型的泛化能力更强，选取了 10 人进行数据集的采集，拍摄场地共有 5 个，减少了环境和人物对于模型的影响．数据集共包括 200 个视频，视频时长为 3～5 s，其示例如图 5 所示．

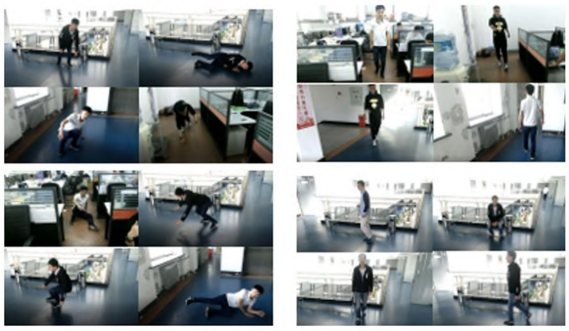

图 5　小样本数据集样例

2.2　空间通道训练方法

本研究对两个通道采取分别训练的策略，如图 6 所示，在空间通道输入静态图像，可以采取迁移学习的方法．但考虑到 ImageNet 数据集的种类远比本研究的动作数据集多，若只是训练最后的全连接层，则会造成卷积层的浪费，效率低而且识别率低，所以本研究只迁移在 ImageNet 上预训练好的模型的前 10 层(第 1 个卷积层到第 3 个 Inception 模块)用于提取边缘、条纹及颜色信息．

先将 UCF101 视频数据集切割为静态图像，作为空间通道训练数据集，而本研究的样本数据集较少，但是与 UCF101 数据集类似，所以可以采用迁移学习的方法．当验证集准确率不再提高时，再对整体进行微调．

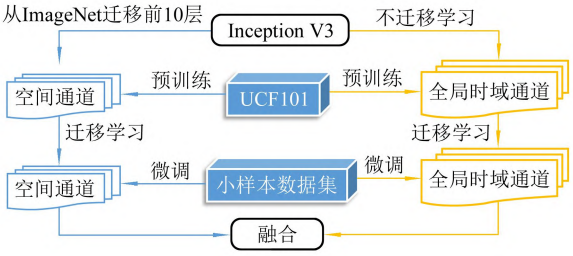

图 6　双通道卷积神经网络的训练

2.3　全局时域通道训练方法

全局时域通道的输入是本文提出的 EMHI 图像，与 ImageNet 数据集完全不同，所以要对其进行整体训练．与空间通道训练类似，先利用 UCF101 视频数据集计算 EMHI 作为全局时域通道数据集，对整个时域通道进行训练，最后迁移学习至本研究中的动作数据集．训练过程如图 6 所示．

3　实验验证

本研究选用 UCF101 数据库对识别效果进行判定，UCF101 数据库包含 101 种动作的 1.332×10^4

段视频，动作的场景复杂．随后将训练好的网络迁移至本研究中的小样本数据集.两个通道单独训练，最后参考文献[1]的方法，对两个通道进行平均融合与支持向量机(SVM)融合．

3.1　空间通道卷积网络

在 UCF101 空间通道数据集上训练至较高的识别率后迁移至小样本数据集进行微调，为了对多帧融合算法的有效性进行探究，在空间卷积通道分别采用 3 帧融合、5 帧融合和 10 帧融合的方式．

测试结果如表 1 所示．对于 UCF101 数据集，文献[1]的空间通道平均识别率为 72.8%，本文方法空间通道平均识别率为 73.1%，利用多帧融合的方式将识别率分别提升到 73.6%，73.9%和 74.2%．在本研究中的小样本数据集上表现更好，空间通道平均识别率为 76.3%，利用多帧融合的方式将识别率分别提升到 76.9%，77.2%和 77.5%．小样本数据集动作类别远少于 UCF101，所以误差更小．而通过多帧融合的方式确实能提高识别准确率，减小误差，证明了该方法的有效性．

表 1　空间通道平均识别

方法	平均识别率/%	
	UCF101	小样本数据集
文献[1]空间通道+softmax	72.8	—
本文空间通道+softmax	73.1	76.3
空间通道+softmax+3 帧融合	73.6	76.9
空间通道+softmax+5 帧融合	73.9	77.2
空间通道+softmax+10 帧融合	74.2	77.5

3.2　全局时域通道卷积网络

利用视频数据集分别计算 MHI 和 EMHI 作为时域通道训练数据集，在 UCF101 时域通道数据集上训练至较高的识别率后迁移至小样本数据集进行微调，分别比较 MHI 和 EMHI 的识别效果．本研究的全局时域通道的输入是单通道的灰度图，而时域通道的输入是 RGB 图．如图 7 所示，本研究在输入层之后多加一层卷积层，卷积核的数量为 3，边界处采取补 0 的方法满足了时域通道的输入层结构．

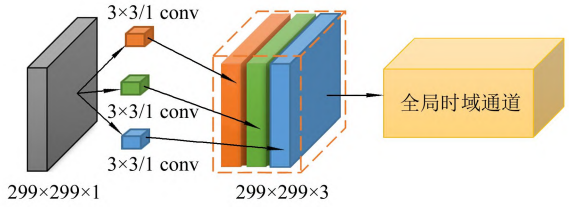

图 7　全局时域通道输入配置

测试结果如表 2 所示．对于 UCF101 数据集，文献[1]的时域通道平均识别率为 81.2%，本文方法利用 MHI 的平均识别率为 77.2%，EMHI 的平均识别率为 82.5%．小样本数据集 MHI 的识别准确率为 81.5%，EMHI 的平均识别率为 87.2%．总体来看，EMHI 的平均识别率要高于 MHI，验证了本文提出的 EMHI 在动作识别中的有效性．

表 2　时域通道平均识别率

方法	平均识别率/%	
	UCF101	小样本数据集
文献[1]时域通道+softmax	81.2	—
MHI+softmax	77.2	81.5
EMHI+softmax	82.5	87.2

3.3　双通道融合

将空间通道卷积网络与全局时域通道卷积网络的识别结果融合，测试方法相同，测试结果如表 3 所示．对于 UCF101 数据集，文献[1]采用平均融合的平均识别率为 86.9%，采用 SVM 融合的平均识别率为 88.0%；本文方法采用平均融合的平均识别率为 88.5%，采用 SVM 融合的平均识别率为 89.7%．在小样本数据集上采用平均融合的平均识别率为 90.3%，采用 SVM 融合的平均识别率为 93.5%．可知空间通道和全局时域通道的深度特征学习能力彼此间互补．

表 3　双通道平均识别率

方法	平均识别率/%	
	UCF101	小样本数据集
文献[1]双通道+平均融合	86.9	—
文献[1]双通道+SVM 融合	88.0	—
本文双通道+平均融合	88.5	90.3
本文双通道+SVM 融合	89.7	93.5

4　结语

本研究提出一种基于空间和全局时域特征的双通道卷积神经网络人体动作识别框架，能够对人体动作信息进行深度特征提取．其中空间通道采用多帧融合的方式进行识别，实验结果表明该方法能有效提高空间通道的识别准确率；时域通道采用基于运动能量的具有自适应能力的 EMHI，相比传统的 MHI 能够更加有效地提取全局动作时域特征．双通道采取融合的方式对动作综合识别，实验结果表明两个通道彼此互补，提高了动作识别的精度．此外本研究利用大型动作数据集进行预训练，迁移至小样本数据集，表现出更好的识别精度，验证了该方法的有效性．

参　考　文　献

[1]　Simonyan K，Zisserman A．Two-stream convolutional

networks for action recognition in videos[J/OL]. [2018-06-01]. https://arxiv.org/abs/1406.2199.

[2] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2015: 4305-4314.

[3] Wang P, Cao Y, Shen C, et al. Temporal pyramid pooling based convolutional neural network for action recogni-recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 99: 1-10.

[4] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition [J/OL]. [2018-06-01]. https//www.computer.org/csdl/proceedings-article/cvpr/2016/8851b933/12OmNzX6cjY.

[5] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]// Proc of European Conference on Computer Vision. Cham: Springer, 2016: 20-36.

[6] Ji S, Yang M, Yu K. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221-231.

[7] Du T, Bourdev L, Fergus R, et al. C3D: generic features for video analysis[J/OL]. [2018-06-01]. http://cn.arxiv.org/abs/1412.0767v1.

[8] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition [J/OL]. [2018-06-01]. http://cn.arxiv.org/abs/1801.07455.

[9] Sun L, Jia K, Yeung D Y, et al. Human action recognition using factorized spatio-temporal convolutional networks [J/OL]. [2018-06-01]. http://cn.arxiv.org/abs/1510.00562.

[10] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. [2018-06-01]. https://arxiv.org/abs/1409.1556.

[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[J/OL]. [2018-06-01]. https://arxiv.org/abs/1409.4842.

[12] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[J/OL]. [2018-06-01]. https://arxiv.org/abs/1512.03385.

[13] Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE, 2009: 248-255.

[14] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[J/OL]. [2018-06-01]. https://arxiv.org/abs/1512.00567.

[15] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(3): 257-267.

[16] Bobick A, Davis J. An Appearance-Based Representation of Action[C]// Proc of International Conference on Pattern Recognition. New York: IEEE, 1996: 307-312.

[17] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization[J/OL]. [2018-06-01]. https://arxiv.org/abs/1506.06579.