基于深度学习的人体行为识别研究综述

袁家政[1]　刘宏哲[2]　徐　成[2]　李扬志[2]　米家辉[2]

1 北京开放大学　北京 100081

2 北京联合大学北京市信息服务工程重点实验室　北京 100101

摘　要　近年来，人体行为识别和深度学习是智能视频分析领域的热点及研究趋势，在视频检索、智能监控以及人机交互等方面具有广泛的应用前景及研究意义。传统的方法目前已难以适应复杂场景下的人体行为识别，且不再满足需求，利用深度学习的方法进行人体行为识别是当前的主流算法，但仍存在困难与挑战。根据人体行为识别的发展历程，首先介绍了早期传统的识别方法，然后重点对深度学习框架下的识别方法进行了分类介绍，其中包括当前常见的卷积神经网络、双流网络、混合网络等。此外还对目前公开的常用行为识别数据集进行了分析，并比对了一些深度学习方法在典型数据集上的性能。最后，对人体行为识别的发展方向进行了探讨。

关键词：人体行为识别；深度学习；神经网络；行为数据集

中图法分类号　TP391.4

Review on Human Behavior Recognition Based on Deep Learning

YUAN Jia-zheng[1], LIU Hong-zhe[2], XU Cheng[2], LI Yang-zhi[2] and MI Jia-hui[2]

1 Beijing Open University, Beijing 100081, China

2 Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China

Abstract　In recent years, human behavior recognition and deep learning are hot spots and research trends in the field of intelligent video analysis, which have a wide range of application prospects and research significance in video retrieval, intelligent monitoring and human-computer interaction. The traditional method is difficult to adapt to the human behavior recognition in complex scenes and it can no longer meet the needs. The deep learning method for human behavior recognition is the current mainstream algorithm, but there are still difficulties and challenges. According to the development of human behavior recognition, this paper first introduces early traditional recognition methods and then focuses on the classification of recognition methods under the framework of deep learning, including convolutional neural network, dual flow network, hybrid network and so on. In addition, the commonly used behavior recognition data sets are analyzed and performances of some deep learning methods on typical data sets are compared. Finally, the development direction of human behavior recognition is discussed.

Keywords　Human behavior recognition, Deep learning, Neural network, Behavior dataset

基金项目：国家自然科学基金(61871028, 61871039, 61802019, 61906017)；北京市属高校高水平教师队伍建设支持计划项目(IDHT20170511)；北京联合大学领军人才项目(BPHR2019AZ01)；北京市教委项目(KM201911417001, KZ201951160050)；北京联合大学研究生科研创新资助项目(YZ2020K001)

This work was supported by the National Natural Science Foundation of China (61871028, 61871039, 61802019, 61906017), Supporting Project for the Construction of High-level Teachers in Beijing-affiliated Universities of China (IDHT20170511), Beijing Union University Leading Talent Program of China (BPHR2019AZ01), Beijing Municipal Education Commission Project of China (KM201911417001, KZ201951160050), Graduate Research and Innovation Funding Project of Beijing Union University (YZ2020K001).

通信作者：袁家政(xxtjiazheng@buu.edu.cn)

1 引言

人体行为识别旨在检测视频中的运动目标，提取其外观和运动信息，并对该信息进行分类和决策，从而识别其行为。基于视频的人体行为识别的主要任务包括识别场景中的人物、时间、地点以及行为，即所谓的“W4 系统”。人体行为识别在人类的生产、生活中的主要应用包括基于互联网视频的内容检索、智能监控等[1]。

基于视频的人体行为识别更加关注人体在视频图像中的变化，根据视频图像中的序列，由算法识别出动作意图。人体行为具有极强的多样性、复杂性，因此准确识别人体行为依然存在巨大的挑战。人体行为识别的流程一般为人体目标检测、提取表征、分类器训练和输出识别结果。人体行为表征是从视频数据中提取的关键信息，是对人体行为进行识别的主要依据，表征选取或提取的优劣直接关系到最终结果的准确性[2]。而后，通过机器学习的方法训练分类器，将识别到的动作表征输入此分类器中，最终得到分类结果。

经过多年的研究与发展，国内外对人体行为的识别研究取得了诸多成果，但人体运动的复杂性以及应用场景的多样性使得识别的精确度和效率在很多场景下达不到实用要求。工作场景中，典型的影响因素包括视频拍摄角度不同、场景的背景发生动态变化、环境光照随时间变化以及人的身高体态衣着多样化等[3]。这些都对运动表征的提取造成了巨大困难，进而影响分析和建模。在过去几十年里，出现了大量的行为识别方法和动作识别数据集。

传统方法主要是手工提取特征的方法[4]。首先对视频帧进行采样，然后对手工提取到的特征进行编码，生成特征向量，接着用特征向量训练分类器，最后由该分类器得到最终的结果。在深度学习未被应用到人体行为识别中之前。国内外学者设计了多种手工特征，如轮廓剪影(Human Silhouette)、时空兴趣点(Space-Time Interest Points)、人体关节点(Human Joint Point)和运动轨迹(Trajectories)等[5]，并围绕这些特征进行了大量的研究和尝试。

虽然手工提取特征的方法取得了一定的成果，但是需要耗费大量的人力，且需要一定的专业基础知识。此外，人工特征不适用于大型人体行为识别数据集。基于深度学习的人体行为识别方法大致可分为 3 类，分别是三维卷积网络、双流网络和混合网络。

本文第 1 节概述人体行为识别研究现状；第 2 节简单回顾基于传统方法的人体行为识别；第 3 节主要介绍基于深度学习的人体行为识别方法；第 4 节对于常用数据集以及数据集上的实验进行介绍与比较；最后对人体行为识别的发展进行探讨。

2 基于传统方法的人体行为识别

基于轮廓剪影的方法通常采用背景剔除来提取人体轮廓剪影等检测人体行为区域，并将提取到的整个区域作为行为表征。这种方法在简单背景下效果较好，容易提取到注意力区域，但在图像噪声、人体遮挡和拍着角度等外界影响下识别精度会大大降低。在复杂场景下，很难提取到准确的轮廓信息，也就无法获得行为表征。典型代表如 Bobick 等[6]提出的用背景去除法获得目标轮廓，然后计算这些图像之间叠加产生的轮廓差建立有运动效果的 MEI(Motion Energy Image)，同时利用时间函数构造运动的 MHI(Motion History Image)。这种方法在简单模式下可以很好地获得注意力区域，但在复杂背景下表现较差。

基于时空兴趣点的采样方法通过探测算子检测视频中的时空兴趣点，并从兴趣点周围提取行为特征。该方法不需要将运动物体分割出来，因此在背景复杂的场景下效果依然不错，但是在遮挡和光线变化较大时效果不佳。Laptev[7]将二维空间兴趣点扩展到三维时空兴趣点，对视频行为检测 Harris3D 兴趣点。Harris 点可以检测在时空上具有显著变化的区域，并且能够自适应地选择兴趣点，最后统计像素直方图形成描述行为的特征向量。

基于人体关节点的提取方法通过识别人体当前姿态进而推测各个躯干的关节点位置，从而获得人体行为表征。该方法与轮廓剪影法的优缺点类似。典型代表有 Fujiyoshi 等[8]提出的提取躯干关键点方法，此方法提取了头部加四肢的 5 个关节点，然后对人体进行识别计算。

基于运动轨迹的提取方法通过追踪人体的运动轨迹来分析行为特征。Wang 等[9]在前人的基础上提出了改进的稠密轨迹方法(Improved Dense Tra-

jectories，IDT）。该方法综合了 HOG，HOF，MBH 等的特征[10]，对轨迹全局施加平滑约束，获得了很好的鲁棒性。在深度学习广泛应用到身体行为识别之前，iDT 是基于手工特征的方法中效果最好、应对场景最丰富的算法。但是由于其训练阶段和应用阶段计算复杂度高，因此该方法速度较慢。

3 基于深度学习的人体行为识别方法

3.1 3D 卷积神经网络

行为识别的方法大多基于图像的平面卷积网络（2D）来对单帧图像的 CNN 进行学习训练，这样会丢失视频中两个连续帧之间的关系，造成视频动作信息流失，发挥不出视频应有的作用。3D 卷积网络是 2D 卷积网络的扩展，其主要使用 3D 卷积来捕获时间信息。3D 卷积由 Ji 等[11]首次提出，并应用于行为识别领域。3D 卷积网络将时域引入了 2D 卷积网络，改善了 2D 无法捕获时域信息的缺点。也就是说，3D 卷积内核用由连续多个帧组成的时空立方体来提取表征。

3D 卷积是对 2D 卷积的扩展，添加了时间维度，从而可以提取视频的时空特征。Ji 等[11]提出了一种新的动作识别模型，通过执行 3D 卷积可从空间和时间维度中提取特征，从而捕获多个相邻帧中编码的运动信息。该模型从输入帧中生成多个信息通道，最终的特征表示组合来自所有通道的信息。Tran 等[12]对提出的 3D 卷积网络进行了扩展，构建出了 C3D(Convolutional 3D)的深层网络结构，并在多数超大型数据集上进行了训练。试验证明，C3D 提取的特征是通用、有效且紧凑的。在此基础上，将残差网络（ResNet）与 C3D 相结合，提出了 Res3D 网络[13]。这个网络进一步提高了性能，运行速度比 C3D 快了 2 倍，且准确率也略有提升。Qiu 等[14]从另一个角度研究了卷积核尺寸设计，提出了伪 3D 网络，并通过将 3D 卷积拆成 2D 卷积和 1D 卷积以及不同的链接方式，验证了该方法的有效性。Diba 等[15]为捕捉长时视频高层语义信息，提出了时域 3D 卷积核，并新增了时域变换层 TTL(Temporal Transition Layer)来替换池化层，该网络为 T3D(Temporal 3D ConvNet)，可实现端到端网络训练。

3D 卷积网络通过 3D 卷积核来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。该结构的最大优势在于其速度，这使得 3D 卷积网络有着很好的应用前景，但 3D 卷积网络的识别精度一般低于双流网络结构。

3.2 双流网络

双流网络模型结构的基本原理为对视频中两帧之间的密集光流进行计算，得到密集光流序列，然后将视频图像和密集光流分别训练成为两个独立的网络，让两个网络分别对动作进行独立判断，最后将结果进行融合，得到最终输出结果。这种网络模型优点是精度较高，但相应的运算量较大，速度慢。

Two-Stream 方法是深度学习在该方向的一大主流方向，最早是 VGG 团队在 NIPS 上提出来的[16]。在这之前也曾有人尝试用深度学习来处理行为识别，如李飞飞团队[17]，通过叠加视频多帧输入到网络中进行学习，但遗憾的是这种方法比手动提取特征更加糟糕。当 Two-Stream CNN 方法被提出来后才意味着深度学习在行为识别中迈出了重大的一步。视频可以分为空间和时间两个部分。空间部分每一帧代表的是空间信息，如目标、场景等，而时间部分是指帧间的运动，包括摄像机的运动或者目标物体的运动信息。因此网络相应地由两个部分组成，分别处理时间和空间两个维度。每个网络都是由 CNN 和最后的 softmax 组成，最后的 softmax 的 fusion 主要考虑了两种方法：平均以及在堆叠的 softmax 上训练一个 SVM。

Carreira 等[18]提出动作在单个帧中可能不明确，然而，现有动作识别数据集的局限性意味着性能最佳的视频架构不会明显偏离单图分析，因为它们依赖于在 ImageNet 上训练的强大图像分类器。Diba 等[19]提出 T3D，一方面由于采用了 3D densenet，区别于之前的 inception 和 Resnet 结构；另一方面，TTL 层，即使用不同尺度的卷积（inception 思想）来捕捉讯息。Qiu 等[20]提出了 P3D 改进 ResNet 内部连接中的卷积形式。Ng 等[21]提出用 LSTM 来进行双流网络时域的融合。

Zhu 等[22]提出的 DTPP 解决了视频级的端到端学习问题，利用时间金字塔池层来聚集由空间和时间线索组成的帧级特征。该模型具有紧凑的视频级

表示，并具有多个时间尺度，以及全局和序列感知。Liu 等[23]提出的 STFN 可以有效地使用时空模块来提取信息并进行双流网络间信息的交互。

TSN(Temporal Segments Networks)[24]是在上述 Two-Stream CNN 基础上的改进网络。目前基于 Two-stream 的方法基本上是以 TSN 作为骨干网络。Two-stream 的方法一个很大的弊端就是不能对长时间的视频进行建模，只能对连续的几帧视频提取 temporal context。为了解决这个问题，TSN 网络提出了一种很有用的方法，先将视频分成 K 个部分，然后从每个部分中随机选出一个短的片段，然后对这个片段应用上述 Two-stream 方法，最后对多个片段上提取到的特征进行融合。

Lan 等[25]改进的地方主要在于 fusion 部分，不同的片段应该有不同的权重，而这部分由网络学习而得，最后由 SVM 分类得到结果。Zhou 等[26]提出的改进关注时序关系推理。对于那些仅靠关键帧(单帧 RGB 图像)无法辨别的动作，如摔倒，可以通过时序推理进行分类。除了两帧之间时序推理，还可以拓展到更多帧之间的时序推理。通过对不同长度视频帧的时序推理，最后进行融合得到结果。该模型建立在 TSN 基础上，在输入的特征图上进行时序推理，增加 3 层全连接层来学习不同长度视频帧的权重。Zhao 等[27]提出 two-in-one，这是一种新 layer 能够结合 RGB 图像与 optical-flow 图像。

双流网络架构虽然精度高，但是由于提取光流的过程非常耗时，因此在现有的硬件条件下无法达到实时检测的要求。

3.3 混合网络

CNN-LSTM[28]是混合网络的代表，其在空间运动模式、时间顺序和长期依赖关系的获取等方面取得了较好的效果。

融合 CNN-LSTM 网络的重点在于提取视频数据中的时空信息。融合 CNN-LSTM 结构可以理解成电路中的串联结构，这种网络结构在早期得到了广泛应用，且识别的精度较高。Karpathy 等[29]研究了时空网络中的几种融合方式，如晚融合、早融合和慢融合，因此能够获取视频中的时序信息。研究结果表明，慢融合比早融合和晚融合的效果更好。Donahue 等[30]提出了 LRCN(Long-term Recurrent Convolutional Network)网络。对视频进行分析处理的关键在于对时序特征的学习和理解，故将 CNN-LSTM 相结合来提取视频数据中的时空信息。

不同于 two-stream 结构和 3D-ConvNet 结构，CNN-LSTM 结构的输入数据模态一般为骨架序列。首先通过 RGB 图像进行关节点估计或利用深度摄像机获取人体骨架序列，每一帧对应人体关节的坐标位置信息，一个时间序列由若干帧组成；然后用 CNN-LSTM 网络对构建出的骨架时序图提取高层特征；最后用 softmax 分类器进行分类。CNN-LSTM 结构的优点在于识别精度高、算法速度较快，是目前行为识别领域最为主流的识别算法。

4 人体行为识别数据集及精度基准

4.1 常用公开数据集

评价不同识别方法的性能，如今有许多公开数据集可以使用，表 1 所列为当前公开的主流数据集。

表 1 常用公开数据集

Table 1 Common public datasets

数据集	年份	视频数量	行为类别	模态
KTH	2004	2391	6	RGB
HMDB51	2011	6766	51	RGB
UCF101	2012	13320	101	RGB

KTH 数据集[31]发布于 2004 年，包含 6 类人体行为：行走、慢跑、奔跑、拳击、挥手和鼓掌，每类行为由 25 个人在 4 种不同的场景(室外、伴有尺度变化的室外、伴有衣着变化的室外、室内)执行多次，相机固定。该数据库总共有 2391 个视频样本。视频帧率为 25 fps，分辨率为 160×120，平均长度为 4 s。

HMDB51 数据集[32]是从各种互联网资源和数字化电影中收集的，其中的人为动作主要是日常行为。该数据集中的一些关键挑战主要是摄像机视点和运动的变化、背景杂乱、志愿者位置和外观的变化。HMDB51 包含 51 个不同的动作类别，每个动作类别包含至少 101 个剪辑，总共 6766 个视频剪辑。此动作类别主要分为 5 种类型：1)一般面部动作；2)交互的面部动作；3)一般身体动作；4)物体交互动作；5)人体交互的身体动作。对于每个动作类别，视频剪辑被分成具有 70 个剪辑的训练集和具有

30 个剪辑的测试集,并且训练集和测试集中的剪辑不能来自同一个视频文件。

UCF-101 数据集[33]来源于国外某视频网站,该数据集是 UCF-50 数据集的扩充,包含 101 个行为,13 320 个视频。101 个动作类别的视频分为 25 个组,每个组可以包含 4～7 个动作的视频,动作类别可以分为 5 种类型:1)人与物体的互动;2)仅身体动作;3)人与人的互动;4)演奏乐器;5)运动。来自同一组的视频可能具有一些共同的特征,如相似的背景,相似的观点等。在动作方面具有最大的多样性,并且在摄像机运动、物体外观和姿势、物体比例、视点、杂乱的背景、照明条件等方面存在很大的差异,是目前为止类别较多、挑战性较大的数据集。

4.2 不同识别方法的性能比较

本节主要比较了当前深度学习的主流方法在不同数据集上的识别精度,如表 2 所列,在 UCF01 和 HMDB51 数据集上对各算法的 mAP 进行了比较。

表 2 不同识别算法的比较

Table 2 Comparison of different recognition algorithms

(单位:%)

方法	UCF101 *mAP*	HMDB51 *mAP*
Two-Stream[34]	94.6	70.3
3D-ConvNet[35]	97.3	78.7
CNN-LSTM[36]	75.8	44.0

表 2 比较了当前较为主流的识别算法的平均精度,所用到的数据模态大多为 RGB 和光流,主要比较了双流网络、3D 卷积网络、CMM-LSTM 3 种方法。结果表明,基于深度学习的行为识别算法精度得到了明显提升。对于算法的模型结构而言,由于 Two-Stream 结构和 3D-ConvNet 结构可以对视频序列获取时序信息,目前这两种模型结构的识别精度较高。

结束语 深度学习是提高识别准确率非常有效的方法,但目前仍然面临一定的挑战。受硬件算力限制,深度学习的行为识别方法只能将少量的连续帧输入网络中以提取特征,而不能将整段视频直接输入网络中进行提取,因此可能出现丢失关键信息或关键动作的缺陷。深度学习的方法均需要大量标注数据进行训练,训练用数据量大,则结果也相对精确。但是在实际应用中,视频拍摄场景、光照、任务等情况千差万别,无法做到每个场景都取得大量数据标注后进行训练。如何做到低成本标注或仅需少量标注甚至无需标注的行为识别算法是未来很有前景的研究方向之一。

参 考 文 献

[1] BREZEALE D,COOK D J. Automatic Video Classification:A Survey of the Literature[J]. IEEE Transactions on Systems Man & Cybernetics Part C,2008,38(3):416-430.

[2] ZHU F,SHAO L,XIE J,et al. From handcrafted to learned representations for human action recognition:A survey [J]. Image Vis. Comput.,2016(55):42-52.

[3] GORELICK L,BLANK M,SHECHTMAN E,et al. Actions as space-time shapes[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2008,29(12):2247-2253.

[4] DALAL N. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. IEEE,2005.

[5] DALAL N,TRIGGS B,SCHMID C. Human Detection Using Oriented Histograms of Flow and Appearance [C] // European Conference on Computer Vision. Springer,Berlin,Heidelberg,2006.

[6] BOBIC K,AARON F. The Recognition of Human Movement Using Temporal Templates [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence,2001.

[7] LAPTEV I. On Space-Time Interest Points[J]. International Journal of Computer Vision,2005,64(2/3):107-123.

[8] FUJIYOSHI H,LIPTON A J,KANADE T. Real-Time Human Motion Analysis By Image Skeletonization[J]. ICE Transactions on Information and Systems,2004,87-D(1):113-120.

[9] WANG H,SCHMID C. Action Recognition with Improved Trajectories[C]// Proceedings of IEEE International Conference on Computer Vision. 2013:3551-3558.

[10] LAPTEV I,MARSZALEK M,SCHMID C,et al. Learning Realistic Human Actions from Movies[C]// Proceedings of IEEE Conference on Computer Vision

and Pattern Recognition. 2008:1-8.

[11] JI S W, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[C]// Proceedings of the International Conference on Machine Learning. 2010:495-502.

[12] TRAN D, BOURDEV L, FERGUS R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// Proceedings of IEEE International Conference on Computer Vision. 2015:4489-4497.

[13] TRAN D, RAY J, SHOU Z, et al. ConvNet Architecture Search for Spatiotemporal Feature Learning[J]. arXiv: 1708.05038, 2017.

[14] QIU Z, YAO T, MEI T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks [C]// International Conference on Computer Vision. 2017:5534-5542.

[15] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification[J]. arXiv:1711.08200, 2017.

[16] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Advances in neural information processing systems. 2014:568-576.

[17] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[18] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6299-6308.

[19] DIBA A, FAYYAZ M, SHARMA V, et al. Temporal 3d convnets: New architecture and transfer learning for video classification[J]. arXiv:1711.08200, 2017.

[20] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3d residual networks[C]// proceedings of the IEEE International Conference on Computer Vision. 2017:5533-5541.

[21] NG Y H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015:4694-4702.

[22] ZHU J, ZHU Z, ZOU W. End-to-end video-level representation learning for action recognition[C]// 2018 24th International Conference on Pattern Recognition (ICPR). IEEE, 2018:645-650.

[23] LIU Z, HU H, ZHANG J. Spatiotemporal Fusion Networks for Video Action Recognition[J]. Neural Processing Letters, 2019:1-14.

[24] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]// European Conference on Computer Vision. Springer, Cham, 2016:20-36.

[25] LAN Z, ZHU Y, HAUPTMANN A G, et al. Deep local video feature for action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017:1-7.

[26] ZHOU B, ANDONIAN A, OLIVA A, et al. Temporal relational reasoning in videos[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018:803-818.

[27] ZHAO J, SNOEK C G M. Dance with Flow: Two-in-One Stream Action Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:9935-9944.

[28] SRIVASTAVA N, MANSIMOV E, SALAKHUDINOV R. Unsupervised learning of video representations using lstms[C]// International conference on machine learning. Boca Raton, 2015:843-852.

[29] KARPATHY A, TODERICI G, SHETTY S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Computer Vision & Pattern Recognition. IEEE, 2014:1725-1732.

[30] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4):677-691.

[31] SCHULDT C, LAPTEV I, CAPUTO B. Recognizing human actions: a local SVM approach[C]// International Conference on Pattern Recognition. IEEE, 2004.

[32] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: A Large Video Database for Human Motion Recognition[C]// IEEE International Conference on Computer Vision. IEEE, 2011:2556-2563.

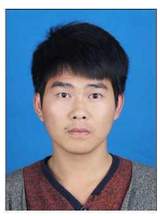
[33] SOOMRO K, ZAMIR A R, SHAN M. UCF101: a dataset of 101 human actions classes from videos in the wild [J]. arXiv:1212, 0402, 2012.

[34] FEICHTENHOFER C,PINZ A,WILDES R P. Spatiotemporal Residual Networks for Video Action Recognition[J]. Neural Information Processing Systems,2017:3468-3476.

[35] TRAN D,WANG H,TORRESANI L,et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:1010-1019.

[36] SRIVASTAVA N,MANSIMOV E,SALAKHUTDINOV R. Unsupervised Learning of Video Representations using LSTMs[C]// International Conference on-Machine Learning. 2015:843-852.

YUAN Jia-zheng, born in 1971, Ph.D. He is now a professor of software engineering of Beijing Open University in Beijing of China. His main research interests include graph and image processing, machine learning and artificial intelligence.



LI Yang-zhi, born in 1996, M. S candidate. His main research interests include computer vision and artificial intelligence.