doi:10.11835/j.issn.1000-582X.2017.11.012

基于深度卷积神经网络和深度视频的人体行为识别

刘 智[a],冯 欣[a],张 杰[b]

(重庆理工大学 a.计算机科学与工程学院;b.电子信息与自动化学院,重庆 400054)

摘要:传统人体行为识别基于人工设计特征方法涉及的环节多,具有时间开销大,算法难以整体调优的缺点。以深度视频为研究对象,构建了 3 维卷积深度神经网络自动学习人体行为的时空特征,使用 Softmax 分类器进行人体行为的分类识别。实验结果表明,提出的方法能够有效提取人体行为的潜在特征,不但在 MSR-Action3D 数据集上能够获得与当前最好方法一致的识别效果,在 UTKinect-Action3D 数据集也能够获得与基准项目相当的识别效果。本方法的优势是不需要人工提取特征,特征提取和分类识别构成一个端到端的完整闭环系统,方法更加简单。同时,研究方法也验证了深度卷积神经网络模型具有良好的泛化性能,使用 MSR-Action3D 数据集训练的模型直接应用于 UTKinect-Action3D 数据集上行为的分类识别,同样获得了良好的识别效果。

关键词:深度学习;人体行为识别;深度卷积神经网络;深度视频;3 维卷积

中图分类号:TP181 **文献标志码**:A **文章编号**:1000-582X(2017)11-099-08

Action recognition based on deep convolution neural network and depth sequences

LIU Zhi[a], *FENG Xin*[a], *ZHANG Jie*[b]

(a. College of Computer Science and Engineering, Chongqing University of Technology; b. College of Electronic Information and Automation, Chongqing University of Technology, Chongqing 400054, P.R.China)

Abstract: Traditional methods for action recognition include several isolated processes and depend on well-designed features, which makes them has the shotcomings of large time cost and difficult to optimize the parameters from the whole. In this paper, we use depth sequences to study deep learning-based action recognition and construct a 3D-based deep convolution neural network to automatically learn spatio-temporal features from raw depth sequences. A Softmax classifier is used on the learned features to take action recognition. Experimental results demonstrate that our method can learn feature representation automatically from depth sequences. The proposed method performs comparable results to the state-of-the-art methods on the MSR-Action3D dataset and achieves good performance in comparison to baseline methods on the UTKinect-Action3D dataset. And the proposed method is simpler in feature extracting and

收稿日期:2016-02-26

基金项目:国家自然科学基金-青年科学基金资助项目(61502065);重庆市教委科学技术研究资助项目(KJ1600937,KJ1500922,KJ1501504)。

Supported by National Natural Science Foundation of China for Young Scientists (61502065) and Scientific and Technological Research Program of Chongqing Municipal Education Commission (KJ1600937, KJ1500922, KJ1501504).

作者简介:刘智(1977—),男,重庆理工大学副教授,博士,主要从事计算机视觉、深度学习、信息融合方向研究,(E-mail) liuzhi@cqut.edu.cn。

action recognition consist of a closed loop system which can learn features automatically. We further investigate the generalization of the trained model by transferring the learned features from one dataset (MSR-Action3D) to another dataset (UTKinect-Action3D) without retraining and obtain very promising classification accuracy.

Keywords: deep learning; human action recognition; deep convolution neural network; depth sequence; 3-dimension convolution;

作为视频分析中的一项流行技术,人体行为识别已逐渐开始应用于日常生活,如自动监控中的异常事件检测,视频检索,人机接口等。传统的人体行为识别包括3个步骤:特征提取,特征表示和识别分类。首先,从视频序列中提取人工特征。如时空兴趣点(STIP, space time interest points)[1],视觉词袋(BOVW, bag of visual words)[2-3],方向梯度直方图(HOG, histograms of oriented gradient)[4-6],和运动历史图像(MHI, motion history image)[7]等都是研究中经常使用到的一些人工特征。其次,使用一些变换和聚类等技术将提取的特征构造出更具区分性的描述子特征,如傅里叶时态变换(FTT, fourier temporal transformation)[8-9]和K均值聚类[10]等。最后,使用分类器对描述子特征进行分类识别,如SVM,Adaboost等。基于特征提取的行为识别方法取得了非常可喜的研究成果。Yang等[4]通过将深度视频投影到3个正交的二维平面空间并逐帧累积整个视频序列的全局行为,提出了深度运动图(DMM, depth motion maps)特征,然后根据DMM计算HOG特征作为每个行为视频的描述子。借用视觉词袋模型的思想,Li等[10]提出了3维点词袋(bag of 3D points)来描述一系列的显著姿势,这些姿势作为节点用于构建表示人体行为的行为图。Roshtkhari等[2]构建了时空视频卷码本,通过将该码本组合成更大的前后文卷,可用于训练视频卷及其时空组合的概率模型。在文献[11]中,Zanfir等运用关节的布局,速度和加速度等信息提出了移动姿势描述子特征。Vemulapalli等[9]基于深度视频的关节骨架信息建模人体不同部位之间的几何信息,从而可将人体行为建模为李群空间[12]的一条弧线。和文献[9]不同,Xia等[13]使用3维关节位置直方图(HOJ3D, histograms of 3D joint locations)来表示不同行为。基于人工特征的行为识别方法近些年的研究进展缓慢,主要是因为,其一,为有效保留行为信息,提取的特征维数越来越高,计算开销太大,很难做到实时性;其二,人工设计特征针对某一具体数据集调校而成,很难泛化到其他数据集;其三,传统行为识别方法各步骤间是孤立的,分类结果好坏并不能自动反馈到特征提取和描述环节。2006年,Hinton等[14]提出了深度学习的概念,其逐层训练算法能够很好训练深度神经网络。从此,深度学习及神经网络又一次得到了研究者的重视并广泛应用于图像分类、语音识别、物体识别等领域[15]。Ji等[15]通过构建3维卷积神经网络(CNN, convolution neural network)模型来研究基于RGB视频的行为识别,他们首先使用一系列的固定核函数为每一帧生成多通道信息,然后通过3维卷积捕捉多个相邻帧之间的运动信息,最后通过组合所有通道的信息得到最终的特征表示。Le等[16]综合了独立子空间分析(ISA, independent subspace analysis)和CNN的优势,首先使用ISA学习行为视频中的不变时空特征,所学到的特征作为多层CNN网络的输入,从而利用CNN学习行为视频的更高层和更抽象特征。Lin等[17]采用文献[16]类似的实验框架,唯一不同的是,在输入CNN前使用ISA,而Lin等[17]则使用慢特征分析(SFA, slow feature analysis)技术进行特征提取。ISA旨在学习空间不变性特征,而SFA更倾向于发现视频序列中的稳定和变化较慢的规律。Du等[18]则分别从视频帧和连续帧之间的光流信息提取空间和时间特征,然后使用2个深度网络进行高层特征的提取并用于人体行为的识别。Tran等[19]以RGB视频为研究对象,直接使用3维卷积提取行为中的特征,取得了较好的效果。现有基于深度学习的行为识别研究,往往是先从视频序列中提取人工设计特征,然后将提取的特征应用于深度神经网络,深度神经网络更多的起到特征降维的作用,其本质上类似于基于人工特征的方法。除此之外,当前的行为识别研究有的基于RGB视频[15-20],有的基于深度视频[21-22]。相对于RGB视频,深度视频中包含了物体的深度信息和几何结构信息,因此它对光线的变化不敏感[23],并且在视频分割、物体检测和行为识别等视觉任务中比RGB视频具有更好的区分性[24]。结合深度视频的优点,以深度视频为研究对象,使用3维卷积构建深度神经网络

模型，直接从行为视频序列自动学习其高层特征表示并进行行为的识别。所提出方法在 UTKinect-Action3D[13] 和 MSR-Action3D[10] 2 个数据集上进行了评估，结果表明，方法在 UTKinect-Action3D 和 MSR-Action3D 数据集均获得了良好的识别性能。

研究的主要贡献如下：

1）提出了 3 维卷积深度神经网络模型用于深度视频序列的行为识别，该模型能对原始深度视频进行自动特征提取并进行分类识别；

2）方法不依赖于复杂的人工设计特征，识别效果在两个常用公开数据库获得了良好的识别性能；

3）探讨了深度神经网络模型的泛化性能，从实验角度验证了深度神经网络模型具有良好的泛化性能。

1　基于 3 维卷积的深度神经网络模型

1.1　网络总体框架

近年来，深度神经网络的研究无论在理论还是实际应用方面均取得了一系列研究成果。根据网络基本结构的不同，深度神经网络可分为深度卷积网络，深度信念网络，堆叠自动编码器等[25]。而在图像处理，行为识别，视频分割，等视觉类任务中，由于深度卷积网络具有建模方便、训练过程简单、识别性能好的特点，因而得到了更多的研究和应用。传统图像处理以二维卷积构造的深度神经网络，不适用于具有 3 维结构的视频。受图像处理中的二维卷积深度神经网络的启发，通过构建 3 维卷积深度神经网络来自动提取行为视频中的空间和时态特征，并用于对人体行为的分类识别。图 1 给出了设计的基于 3 维卷积的深度神经网络模型。该网络具有 2 个 3 维卷积层（convolution layer），其中的卷积操作同时考虑了空间和时间维度，2 个卷积层的特征图数目分别为 32 和 128。由于使用的 2 个数据集的视频大小不一样，因此采用了不同的卷积核大小，对于 MSR-Action3D 数据集，卷积核大小分别为 $5\times5\times7$ 和 $5\times5\times5$，而 UTKinect-Action3D 数据集卷积核大小分别为 $5\times5\times5$ 和 $5\times5\times5$。每个卷积层后是池化层（pooling layer），使用的是最大池化（max pooling）技术，池化操作可以实现对提取特征的平移不变性。卷积层和池化层构成该深度神经网络模型的主体部分。然后是向量化层、2 个全连接层（full connected layer）和分类层，全连接层神经元个数分别为 2 056 和 512，采用的是传统的前馈式神经网络连接方式；分类层中采用的是 Softmax 分类器，网络中的激活函数全部为双曲正切函数 tanh。和一般的深度网络一样，使用反向传播（BP）算法训练基于 3 维卷积的深度神经网络。实验时，采用随机梯度下降法（SGD）进行深度学习，训练时的学习速率和权重衰减系数均为 1×10^{-4}，冲量单元为 0。

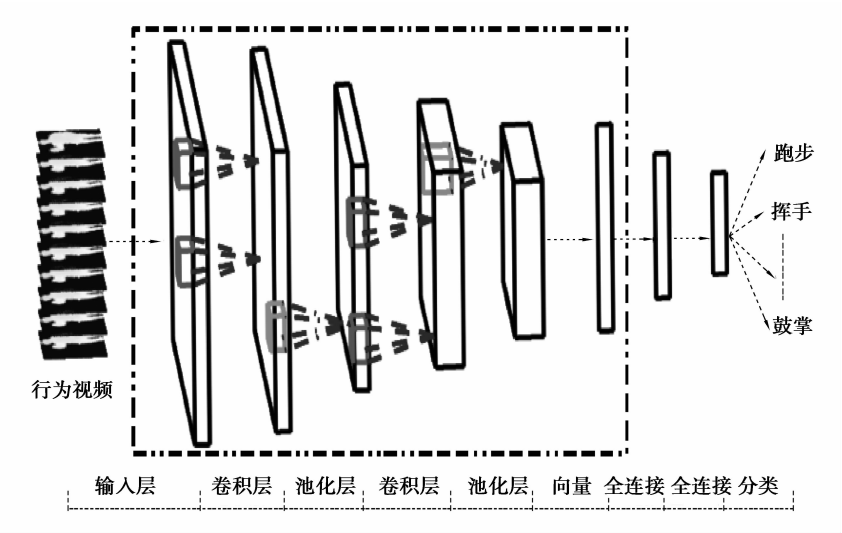

图 1　基于 3 维卷积的深度神经网络模型

Fig.1　The deep neural network model based on 3-dimensional convolution

1.2 3维卷积与池化的数学建模

卷积神经网络有神经认知的根源,其卷积和池化层灵感直接来源于视觉神经科学中的简单细胞和复杂细胞。图像处理领域使用的是二维卷积和池化技术,不适用于3维视频。通过构建基于3维卷积和池化的深度神经网络进行人体行为的识别。

1.2.1 3维卷积运算

假定卷积前后分别为$l-1$和l层,特征图数分别为N和M,卷积前视频大小为$T\times W\times H$,T,W,H分别为视频的帧数、帧宽和帧高,各维度卷积步长均为1,卷积核大小为$k_T\times k_W\times k_H$。由于使用不同卷积核生成不同特征图过程是相同的,后续公式只考虑一个特征图。则卷积后l层大小为$(T-k_T+1)\times(W-k_W+1)\times(H-k_H+1)$,获取位置$(t_l,i_l,j_l)$激活值的3维卷积操作运算定义如公式(1)和公式(2)所示

$$x_{t_l i_l j_l}=\sum_{k=1}^{N}\Big(\sum_{t_{l-1}=t_l}^{t_l+k_T-1}\sum_{i_{l-1}=i_l}^{i_l+k_W-1}\sum_{j_{l-1}=j_l}^{j_l+k_H-1}w_{t_{l-1}i_{l-1}j_{l-1}}^{k}x_{t_{l-1}i_{l-1}j_{l-1}}^{k}+b^k\Big),\tag{1}$$

式中:$x_{t_l i_l j_l}$表示第l层位置(t_l,i_l,j_l)的输入加权和(包括偏置单元);$t_l\in[1,T-k_T+1]$;$i_l\in[1,W-k_W+1]$;$j_l\in[1,H-k_H+1]$,k为$l-1$层特征图编号

$$a_{t_l i_l j_l}=f(x_{t_l i_l j_l}),\tag{2}$$

式中:$a_{t_l i_l j_l}$表示第l层位置(t_l,i_l,j_l)的激活值(输出值);激活函数$f(\cdot)$为双曲正切函数,如公式(3)所示。

$$\tan h(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}。\tag{3}$$

1.2.2 三维池化运算

基于三维卷积的深度神经网络中每个卷积层后接一个池化层,池化技术能提高算法的平移不变性,本研究使用的是最大池化技术,池化区域大小为$2\times2\times2$,各维度池化步长均为2,则最大池化如公式(4)所示。

$$x_{t_{lp}i_{lp}j_{lp}}=\max(x_{t_l i_l j_l}),\tag{4}$$

其中:$t_l\in[2t_{lp}-1,2t_{lp}+1]$;$i_l\in[2i_{lp}-1,2i_{lp}+1]$;$j_l\in[2j_{lp}-1,2j_{lp}+1]$;$x_{t_{lp}i_{lp}j_{lp}}$为$l$层使用最大池化后,位置$(t_{lp},i_{lp},j_{lp})$的特征值,即$l$层该特征图在位置$(t_{lp},i_{lp},j_{lp})$的特征值。

2 实验数据及结果讨论

2.1 实验设置及预处理

使用UTKinect-Action3D[13]和MSR-Action3D[10]2个公开数据集来评价所提出方法,它们均是使用深度摄像机(Kinect)拍摄的人体行为视频数据。UTKinect-Action3D中共有10个行为,分别是Walk,Sit down,Stand up,Pick up,Carry,Throw,Push,Pull,Wave hands和Clap hands。共有10个被试,每个被试拍摄每个行为2次,由于第10个被试Carry行为的第2次拍摄被认为是无效视频,因而总共有199个有效视频。为了计算方便,使用了全部200个视频。MSR-Action3D数据集中有20个行为,由10个被试拍摄完成,其中每个被试完成每个行为2~3次。根据该视频数据集基础项目研究中的实验设置[10],20个行为分为3个行为子集,分别是AS1,AS2和AS3(如表1所示),每个行为子集中包含8个不同的行为。为降低对不同实验结果的影响,在实验前对每个视频进行简单的预处理(见图2):1)背景去除:在深度视频中,背景的深度信息是一致的,而前景的深度信息是有变化的,可根据该特点去除背景信息;2)边界框确定:针对每一个视频,分别根据其每一帧,得出能并且仅能框住人体行为的边界框,取所有帧的最大边界框作为本视频的边界框(见图2);3)规范化:使用插值技术将上一步处理后的所有视频规范化到统一大小,其中规范化后的视频帧数等于所有视频帧数的中间值。同时使用min-max方法将将所有视频的深度信息值规范化到[0,1]范围;最后,将所有样本进行水平翻转形成新的样本从而成倍扩大数据集中的训练样本。预处理后,UTKinect-Action3D和MSR-Action3D的行为视频大小分别为$28\times32\times32$和$38\times32\times32$,其中从前往后依次为视频中的帧数,帧宽和帧高。实验深度神经网络模型部分采用Torch平台[26]进行编写,数据预处理部分则使用Matlab平台完成。

表 1　MSR-Action3D 数据集中的行为子集 AS1,AS2 和 AS3

Table 1　The action subset AS1,AS2 and AS3 in MSR-Action3D dataset

AS1	AS2	AS3
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

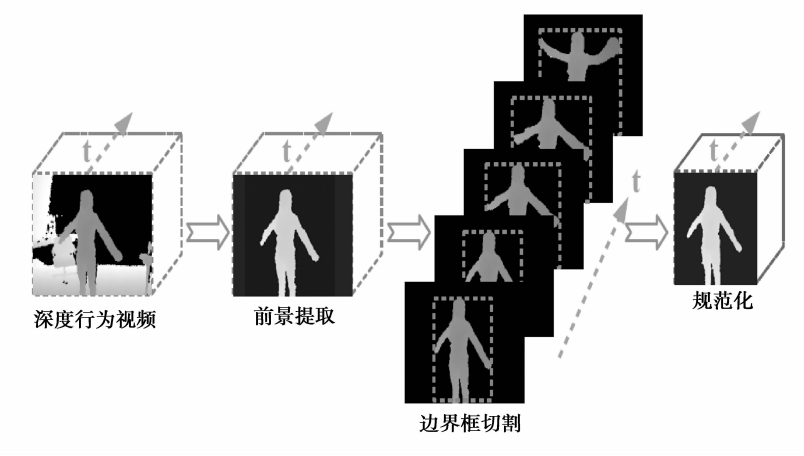

图 2　数据预处理简要步骤图

Fig.2　The steps of data preprocessing

2.2　MSR-Action3D 数据集上的识别性能

首先在 MSR-Action3D 数据集上验证了研究方法的有效性,按照文献[10]的实验设置,研究方法和该数据集的基准项目研究[10]及近些年基于人工特征提取的几个主要方法进行了比较。表 2 给出了方法和在 3 个不同行为子集上的行为识别准确度。从识别结果可以看出,基于 3 维卷积深度神经网络的人体行为识别方法能有效对人体行为进行识别,各行为子集识别准确度和平均准确度均要优于该数据集的基准项目研究。其主要是因为使用 3 维词袋模型提取行为视频中的特征,该特征能提取视频中有代表性的 3 维词袋信息,但忽略掉了视频中空间和时态信息,而基于 3 维卷积深度神经网络的人体行为识别方法对视频采用 3 维卷积操作,有效的保持了空间和时态特征,因而获得了更好的性能。表 3 给出了研究方法与当前识别效率最好的方法文献[8]进行了比较,该实验采用文献[8]中的方法的设置。结果表明研究提出的方法能与文献[8]中方法保持一致的识别性能,充分说明了本方法的有效性。

表 2　研究与 MSR-Action3D 数据集基准研究项目的比较

Table 2　Comparison with the benchmark project in MSR-Action3D dataset　%

方法	AS1	AS2	AS3	平均
文献[10]	72.9	71.9	79.2	74.7
研究方法	84.72	78.95	88.16	83.94

表3 研究与文献[8],文献[10]在AS3上的识别性能比较

Table 3 Performance evaluation compared with [8]and [10] %

方法	识别准确率
文献[10]	79.20
研究方法	88.16
Actionlet 文献[8]	88.20

2.3 UTKinect-Action3D 数据集上的识别性能

在 UTKinect-Action3D 数据集,研究方法与该数据集上的基准研究项目进行了比较[13]。文献[13]使用 Leave-One-Out 交叉验证方法(LOO-CV)。为实验的方便,研究使用 Leave-One Subject-Out 交叉验证(LOSO-CV),即每次只将一个被试的所有行为视频作为测试集,而其他被试的数据作为训练集,从而为每个被试训练出一个深度神经网络模型,显然该实验条件比文献[13]更为苛刻。表4给出了研究方法在不同被试上的行为识别准确度。从表4可以看出,各被试行为识别准确率平均值为82%,基本能正确识别绝大部分被试的行为,而被试5,6,7,10上的识别准确率相对较低,可能是因为 UTKinect-Action3D 是个多视角数据集,这几个被试在进行动作行为拍摄时视角的偏差过大所致。然而,研究方法与文献[13]90.92%的识别率还是有一定差距,相差近8个百分点,可能是因为充分利用了深度视频中的骨架信息,并使用隐马尔科夫模型(HMM, hidden markov model)建立骨架信息的时态模型[13],其缺点是行为识别框架过于复杂,系统性能受骨架信息提取,HOJ3D 特征提取,特征 LDA 投影,行为词聚类和 HMM 模型训练等多个环节的影响,而且文献[23,27]指出提取骨架是个复杂的过程,所提取的骨架信息的准确性取决于深度视频的拍摄情况。同时,研究方法实验条件比文献[13]更为苛刻,而且实验数据相对较小,模型的训练不足,这些也是导致识别效果不够好的原因。尽管如此,相对于文献[13]等工特征提取的方法,基于深度学习的方法具有更好的泛化性能[28],而且研究方法不需要复杂的人工特征提取环节,只需对原始视频进行简单处理,即可由深度神经网络模型进行特征的自动提取并完成识别分类过程,方法简单、涉及环节更少。

表4 UTKinect-Action3D 中各被试行为识别准确率(平均:82%)

Table 4 The action recognition accuracy of each subject in UTKinect-Action3D dataset

被试	被试1	被试2	被试3	被试4	被试5
识别率/%	90	85	90	85	75
被试	被试6	被试7	被试8	被试9	被试10
识别率/%	70	75	90	90	70

2.4 深度卷积神经网络的泛化性能

泛化性能差是人工特征提取的行为识别方法一大缺点,该类方法所提取特征具有数据针对性,在某个数据集上表现良好的特征可能在另一数据集上的性能会急剧下降。图像分类识别上的研究表明,深度神经网络具有良好的泛化性能。Oquab 等先用大数据集训练模型,然后使用目标数据集上训练数据进行模型参数微调,微调后的神经网络模型能在目标数据测试集上获得良好的分类性能[28]。本研究通过简单实验测试了基于三维卷积深度神经网络的泛化性能,将 MSR-Action3D 数据集上训练好的深度神经网络模型,不经过微调,直接用于 UTKinect-Action3D 数据集上的行为分类识别,其识别性能仍然达到了73%的识别准确度,充分表明了深度神经网络在行为识别领域的良好泛化性,为少样本数据集的分类识别带来了曙光。

3 结　语

近些年来，深度学习方法由于设计思想简单，识别效果好，因而在计算机视觉类任务中越来越多的得到研究者的关注。受到基于二维卷积的深度神经网络在图像检测、识别中成功应用的启发，以深度视频为研究对象，通过构建基于三维卷积的深度神经网络来自动学习人体行为的时态和空间特征，并用于人体行为的识别。MSR-Action3D 和 UTKinect-Action3D 数据集上的实验结果表明，研究构建的基于 3 维卷积的深度神经网络模型能对视频中的人体行为进行有效的识别，识别性能与当前主流的方法具有可比性。除此之外，相对于传统基于人工特征的行为识别方法，基于深度卷积神经网络的方法涉及环节少，能自动提取特征，不需要对原始视频进行复杂处理，方法更为简单，更重要的是所提取的特征具有更好的泛化性能，在一个数据集上训练好的模型能够直接应用于其他数据集的分类识别。

参考文献：

[1] Peng X, Qiao Y, Peng Q. Motion boundary based sampling and 3d co-occurrence descriptors for action recognition[J]. Image and Vision Computing, 2014, 32(9): 616-628.

[2] Roshtkhari M J, Levine M D. Human activity recognition in videos using a single example[J]. Image and Vision Computing, 2013, 31(11): 864-876.

[3] O'Hara S, Lui Y M, Draper B A. Using a product manifold distance for unsupervised action recognition[J]. Image and Vision Computing, 2012, 30(3): 206-216.

[4] Yang X, Zhang C, Tian Y L. Recognizing actions using depth motion maps-based histograms of oriented gradients[C]// ACM International Conference on Multimedia. NewYork: ACM, 2012: 1057-1060.

[5] Oreifej O, Liu Z. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences[C]// Computer Vision and Pattern Recognition. NewYork: IEEE, 2013: 716-723.

[6] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]// Computer Vision and Pattern Recognition, 2008, IEEE Conference on. NewYork: IEEE, 2008: 1-8.

[7] Davis J W, Bobick A F. The representation and recognition of human movement using temporal templates[C]// Conference on Computer Vision and Pattern Recognition. Washington, D.C: IEEE Computer Society, 1997: 928.

[8] Wu Y. Mining actionlet ensemble for action recognition with depth cameras[C]// IEEE Conference on Computer Vision and Pattern Recognition. Washington, D.C: IEEE Computer Society, 2012: 1290-1297.

[9] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3D skeletons as points in a lie group[C]// Computer Vision and Pattern Recognition. NewYork: IEEE, 2014: 588-595.

[10] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3D points[C]// Computer Vision and Pattern Recognition Workshops. NewYork: IEEE, 2010: 9-14.

[11] Zanfir M, Leordeanu M, Sminchisescu C. The moving pose: an efficient 3D kinematics descriptor for low-latency action recognition and detection[C]// IEEE International Conference on Computer Vision. NewYork: IEEE, 2014: 2752-2759.

[12] Richard M, Sastry, et al. A mathematical introduction to robotic manipulation[J]. CRC Press, 1994, 39(9): 292.

[13] Xia L, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3D joints[C]// Computer Vision and Pattern Recognition Workshops. NewYork: IEEE, 2012: 20-27.

[14] Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[15] Shuiwang J V, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[16] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis[C]// IEEE Conference on Computer Vision and Pattern Recognition. Washington, D.C: IEEE Computer Society, 2011: 3361-3368.

[17] Sun L, Jia K, Chan T H, et al. DL-SFA: deeply-learned slow feature analysis for action recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Washington, D.C: IEEE Computer Society, 2014: 2625-2632.

[18] Du T, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks[C]// IEEE International Conference on Computer Vision. NewYork: IEEE, 2015: 4489-4497.

[19] Simonyan K, Zisserman A. Two-Stream convolutional networks for action recognition in videos[J]. Advances in Neural Information Processing Systems, 2014, 1(4): 568-576.

[20] Krizhevsky A, Sutskever I, Hinton G E, et al. Imagenet classification with deep convolutional neural networks[J]. Advances in Neural Information Processing Systems, 2012: 1097-1105.

[21] Ye M, Zhang Q, Wang L, et al. A survey on human motion analysis from depth data[M]. Berlin Heidelberg: Springer, 2013: 149-187.

[22] Valle E A, Starostenko O. Recognition of human walking/running actions based on neural network[C]// International Conference on Electrical Engineering, Computing Science and Automatic Control. NewYork: IEEE, 2013: 239-244.

[23] Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from single depth images[J]. Communications of the ACM, 2013, 56(1): 116-124.

[24] Yang X, Tian Y L. Super normal vector for activity recognition using depth sequences[C]//Computer Vision and Pattern Recognition. NewYork: IEEE, 2014: 804-811.

[25] Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives[J]. IEEE Transactions onPattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[26] Collobert R, Kavukcuoglu K, Farabet C, et al. Torch7: a Matlab-like environment for machine learning[C]// BigLearn, NIPS Workshop. Boston: MIT press, 2012.

[27] Shotton J, Sharp T, Kipman A, et al. Real-time human pose recognition in parts from single depth images[J]. Communications of the ACM, 2013, 56(1): 116-124.

[28] Oquab M, Bottou L, Laptev I, et al. Learning and transferring mid-level image representations using convolutional neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Washington, D.C: IEEE Computer Society, 2014: 1717-1724.

（编辑 侯 湘）