文章编号 1004-924X(2017)03-0799-07

构建多尺度深度卷积神经网络行为识别模型

刘 智[1],黄江涛[2*],冯 欣[1]
(1. 重庆理工大学 计算机学院,重庆 400054;
2. 广西师范学院 计算机与信息工程学院,广西 南宁 530001)

摘要:为了减化传统人体行为识别方法中的特征提取过程,提高所提取特征的泛化性能,本文提出了一种基于深度卷积神经网络和多尺度信息的人体行为识别方法。该方法以深度视频为研究对象,通过构建基于卷积神经网络的深度结构,并融合粗粒度的全局行为模式与细粒度的局部手部动作等多尺度信息来研究人体行为的识别。MSRDailyActivity3D 数据集上的实验得出该数据集上第 11~16 种行为的平均识别准确率为 98%,所有行为的平均识别准确率为 60.625%。结果表明,本方法能对人体行为进行有效识别,基本能准确识别运动较为明显的人体行为,对仅有手部局部运动的行为的识别准确率有所下降。

关 键 词:卷积神经网络;深度学习;人体行为识别;计算机视觉;多尺度
中图分类号:TP394.1;TH691.9 **文献标识码**:A **doi**:10.3788/OPE.20172503.0799

Action recognition model construction based on multi-scale deep convolution neural network

LIU Zhi[1], HUANG Jiang-tao[2*], FENG Xin[1]

(1. *College of Computer Science and Engineering, Chongqing University of Technology, Chongqing* 400054, *China*;
2. *College of Computer and Information Engineering, Guangxi Teachers Education University, Nanning* 530001, *China*)
* *Corresponding author, E-mail: hjt@gxtc.edu.cn*

Abstract: In order to simplify the feature extracting process of Human Activity Recognition (HAR) and improve the generalization of extracted feature, an algorithm based on multi-scale deep convolution neural network was proposed. In this algorithm, the depth video was selected as research object and a parallel CNN (Convolution Neural Network) based deep network was constructed to process coarse global information of the action and fine-grained local information of hand part simultaneously. Experiments were executed on MSRDailyActivity3D dataset. The average recognition accuracy on actions ranging from No. 11 to No. 16 was 98%, while that on all actions was 60.625%. The experimental results showed that proposed algorithm could take effective recognition for human activity. Almost all of the actions with obvious movements and most of actions with local movements

收稿日期:2016-12-21;**修订日期**:2017-01-15.
基金项目:重庆市教委科学技术研究基金资助项目(No. KJ1400926);广西自然科学基金重点项目(No. 2014GXNSFDA118037)

just in hands could be recognized effectively.

Key words: convolution neural network; deep learning; human activity recognition; computer vision; multi-scale

1 引　言

目前,有关人体行为识别的研究越来越引起计算机视觉研究工作者的重视,并已广泛应用于自动监控,事件检测,人机接口,视频获取等各个领域。传统的人体行为识别方法主要基于人工设计特征,如方向梯度直方图(Histograms of Oriented Gradient, HOG)[1],运动历史图像(Motion History Image, MHI)[2]等,然后采用支持向量机[3]等分类器对提取的特征进行分类识别。WanqingLi 等人[4]通过提取视频中有代表性的 3D 词袋(Bag of 3D Points,BOPs)来表示人体的一系列姿势,然后以 BOPs 为点构建人体行为图,通过计算行为图上每一条路径的概率进行人体行为识别。文献[2]研究了运动背景下的行为识别,首先提取人体的 MHI 特征,然后用 HOG 进行特征描述,最后使用高斯混合模型(Gaussian Mixture Model, GMM)进行行为的分类识别。Jiang Wang[5]等人则利用深度视频中的骨架信息,通过逐帧计算每个关节相对其他关节的位置和每个关节的局部占位模式(Local Occupancy Patterns,LOP),提出了 actionlet 组合模型来描述人体行为。Lu Xia 和 J. K. Aggarwal[6]先抽取深度视频的时空兴趣点(Spatio-Temporal Interest Points,STIPs),然后以各 STIP 为中心,构造出表示人体行为的深度立方相似特征(Depth Cuboid Similarity Feature, DCSF)。受 HOG 思想的启发,Omar Oreifej 和 Zicheng Liu[7]针对深度视频设计了方向四维法线直方图(Histogram of Oriented 4D Normals, HON4D)特征。为了同时强调人体轮廓和运动的作用,Chenyang Zhang 和 Yingli Tian[8]则对深度运动图(Depth Motion Map, DMM)特征进行扩展,提出了边加强 DMM(Edge Enhanced DMM, E^2DMM)特征。

基于人工特征提取的人体行为识别的研究取得了很多优秀成果[9],然而也存在一些难以解决的问题:提取的特征对训练数据具有依赖性,不易泛化到其他数据;计算开销太大,很难做到实时性。深度学习能自动提取隐藏在数据间的多层特征表示,已经成功应用于语音识别,图像识别与分类,分割等领域。鉴于深度学习的上述优点,Quoc V. Le 等人[10]运用独立子空间分析(Independent Subspace Analysis,ISA)算法自动学习视频数据中稳定的时空特征,然后使用深度结构学习 ISA 的多层表示。文献[11]利用 CNN 构造多层深度结构,提出了 PANDA 算法,用于识别人的属性(如性别、发型、表情等)。DeepPose[12]方法也是基于 CNN 构建深度神经网络,该方法不但用于图像中人体姿势的识别,也对图像中的目标定位进行了探索。文献[13]则基于限制波尔茨曼机,构造出自举深度信念网络,用于人脸的识别。Kaiming He 等人[14]在其最新的研究中同样使用了基于 CNN 的深度神经网络,其贡献在于使用空间池化技术对输入进行处理,从而使得该算法能对任何大小的图像进行分类,而传统基于 CNN 的深度学习方法需要将输入规范化到统一尺寸。为了提高深度学习算法的泛化性能,Min Lin 等人[15]提出了网络嵌套的思想,即网络中的某一个节点可以嵌套一个网络进行学习。文献[16]不但深刻剖析了基于 CNN 的深度神经网络的思想,而且还借鉴了 Min Lin 等人[15]的思想,提出了一个更深层次的网络,取得了较好的效果。

综上,基于特征提取的算法时间开销太大,难以实现实时处理。近些年来,基于 CNN 的深度神经网络在人工智能领域的应用较为广泛,然而关于它的研究主要集中在图像识别、分割、定位等方面,对基于视频的人体行为识别的研究仍比较少。同时相较于传统 RGB 视频,深度视频能提供人体的三维几何信息,而且对光线变化不敏感[17]。基于此,本文以深度视频数据为研究对象,通过构建基于 CNN 的深度神经网络结构,并融合全局的人体行为信息和局部的手部动作等多尺度信息,使用传统的二维 CNN 来研究三维的人体行为识别。本文的创新在于:

(1)使用图像处理中的二维 CNN 构建深度

卷积神经网络并用于人体行为识别；

(2)所提出的方法不依赖于人工设计特征，不需要对数据进行复杂预处理，流程简单。

2　基于多尺度信息融合和深度学习的人体行为识别

传统的基于 CNN 和深度学习的网络结构适合于二维图像的处理，不能直接应用于三维的视频数据集。如果将深度视频的每一帧看做图像的一个特征图(Feature Map，FM)，则一个具有 N 帧的深度视频可以看做具有 N 个 FM 的图像。然而由于描述人体行为的视频帧数 N 一般都比较大，因而直接使用传统网络将带来巨大的时间开销。本文通过构建多个深度网络，组成并行结构来研究深度视频的人体行为识别。首先将深度视频先拆分成多个视频段，然后分别使用各并行分支网络进行学习，再对各网络分支学习到的高层表示进行融合连接，最后将融合后的高层表示送入全连接层和分类层进行分类识别。与此同时，针对 MSRDailyActivity3D 数据集中大部分行为的细微差别主要集中于左手这一特点，如读书、写字、用笔记本电脑、玩游戏等行为。本文除了提取粗粒度的全局行为信息之外，还提取了每个视频左手处的细粒度信息，通过融合粗粒度和细粒度等多尺度信息来完成人体行为识别。对其他数据集则根据具体情况提取不同部位或多个部位的细粒度信息。图 1 给出了具有 3 个 CNN 层和 2 个全连接层的深度网络结构图。根据实验目的的不同，本文使用了不同层数的网络，如表 1 所示。

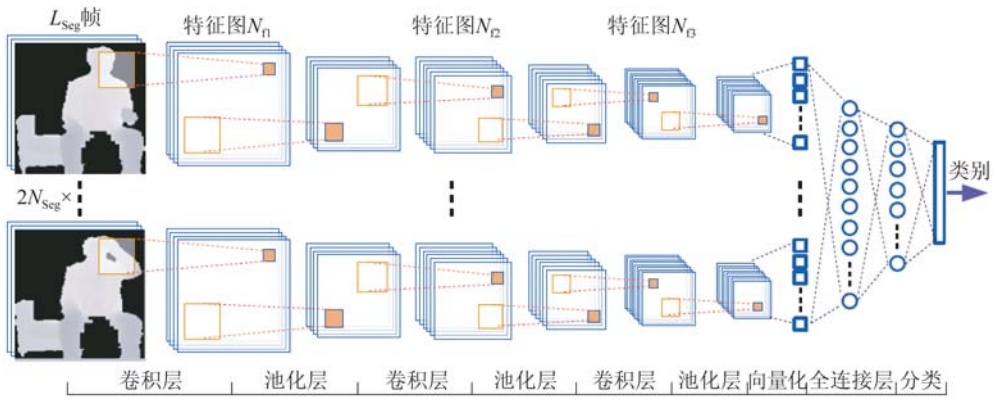

图 1　基于 CNN 和深度学习的人体行为识别框架

Fig. 1　Framework for HAR based on CNN and deep learning

表 1　本文使用到的深度网络及其参数

Tab. 1　Deep networks and their parameters in this paper

网络	层数	卷积核	N_{f1}，N_{f2}，…	全连接层
2CNN2F	2	5×5，5×5	32，128	1 024，512
3CNN2F	3	5×5，7×7，5×5	32，64，128	1 024，512
4CNN2F	4	5×5，5×5，6×6，5×5	16，32，64，128	1 024，512

算法步骤描述：假设规范化后表示一个行为的视频大小为 $N_F\times W\times H$(本文中为 192×128×128)，其中 W，H 分别为视频帧的宽和高。

(1)将帧数为 N_F 的行为视频以 L_{Stride} 为步长进行分段，其中每段包含 L_{Seg} 帧，则分段数为 $N_{Seg}=1+(N_F-L_{Seg})/L_{Stride}$，然后将视频帧 1/4 下采样，则分段后形成了 $N_{Seg}\times L_{Seg}\times W/4\times H/4$ 的视频段矩阵；

(2)以深度视频每一帧的左手关节为中心，截取 $W/4\times H/4$ 大小的帧组成 $N_F\times W/4\times H/4$ 的新视频，对新视频采取步骤(1)方法得到 $N_{Seg}\times L_{Seg}\times W/4\times H/4$ 的视频段矩阵。

(3)将步骤(1)和步骤(2)得到的视频段矩阵进行融合得到 $2N_{Seg}\times L_{Seg}\times W/4\times H/4$ 的视频段矩阵；该视频段矩阵即为深度网络的输入，即该网络具有 $2N_{Seg}$ 个并行深度神经网络，每个深度神经网络的输入为 $L_{Seg}\times W/4\times H/4$ 的视频；

(4)使用训练数据集对并行深度神经网络进行训练，然后使用测试数据集进行人体行为识别的测试，训练数据集和被试数据集完全不相交。

本文中选择{1,3,5,7,9}表演的行为视频用于训练,而将{2,4,6,8,10}表演的行为视频用于测试。

假设 $N_F=192$, $L_{Seg}=16$, $L_{Stride}=16$,则深度神经网络框架需要采用 24 个并行网络,每个网络的输入为 16×32×32 的视频段序列,即每个视频段含有 16 帧视频,视频图像大小为 32×32。

3　实验及讨论

3.1　数据集及预处理

本文使用 Kinect 设备采集的 MSRDailyActivity3D 数据集进行实验[5],该数据集收集了日常生活中常见的 16 种行为:喝水、吃零食、读书、打电话、写字、用笔记本电脑、用吸尘器、欢呼、静止站立、撕纸、玩游戏、躺下沙发、行走、弹吉他、站起和坐下。每个行为动作由同一主试以两种不同的方式完成:坐在沙发上或站着。整个数据集共有 320 个行为视频。图 2 给出了该数据集中的一些行为样例。该数据集记录了人体行为和周围环境的交互,提取出的深度信息含有大量的噪声,而且数据集中的大部分行为只在局部存在细微差异,如图 2,图 3 所示,因而极具挑战性。

在实验前,对每个视频进行简单的预处理,如图 4 所示。(1)背景去除:深度摄像机记录的是每一点的位置信息,相对于运动目标,深度视频中背景的位置信息是固定不变的,根据该特点可去除背景信息;(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,如图 4 所示;(3)规范化,包括空间、时间和深度信息规范化:空间规范化,直接使用 matlab 中的 imresize 函数将图像缩放到指定大小,时间规范化,使用插值技术(公式 1)将所有视频规范化到统一长度,规范化后的视频帧数等于所有视频帧数的中间值,深度信息规范是使用 MinMax 算法将所有视频的像素值规范化到[0,1]范围;(4)将所有样本进行水平翻转形成新的样本,从而使数据集中的训练样本成倍扩大。本文算法采用 Torch 平台[18]进行编写,其中的学习速率为 1×10^{-4},损失函数为平台自带的 Softmax 回归,激活函数为双曲正切(tanh)函数。

$$i=\left\lceil \frac{N}{N_F} * j \right\rceil, \tag{1}$$

其中 N 和 N_F 分别为规范化前后视频含有的帧数,则规范化后第 j 帧来自于规范化前视频中的第 i 帧。其中的括号为上取整。

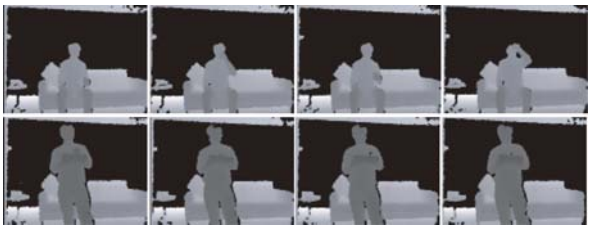

图 2　MSRDailyActivity3D 中的行为视频
(处理前,上:喝水,下:写字)

Fig. 2　Activity videos before processing
(top: drinking, bottom: writing)

图 3　MSRDailyActivity3D 中的行为视频
(处理后,上:喝水,下:写字)

Fig. 3　Activity videos after processing
(top: drinking, bottom: writing)

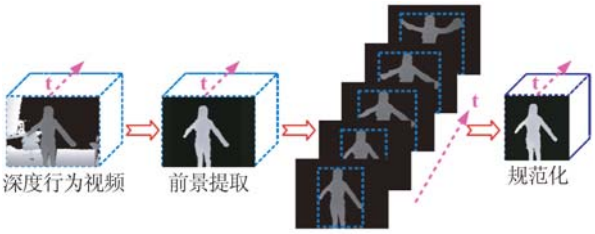

图 4　行为视频预处理简要步骤

Fig. 4　Brief steps of retreatment for activity videos

3.2　基于多尺度信息融合和深度学习的 HAR 识别

根据第 2 节的描述,本文使用表 1 中的 2CNN2F 网络,将粗粒度的全局行为识别视频和细粒度的手部动作序列等多尺度信息作为深度网络的输入。本节实验中的 L_{Stride} 和 L_{Seg} 均设置为 16,即将抽取整个视频的 12×16×32×32 的全局行为序列和 12×16×32×32 局部手部动作序列合并,形成 24×16×32×32 输入视频矩阵。表 2 给出了本文方法与其他方法在 MSRDailyActivity3D

数据集上识别性能的对比结果。其中 2CNN2F 是指仅使用粗粒度的全局行为信息，而 2CNN2F_J 则表示多尺度信息融合方法，它融合了粗粒度的行为模式信息和细粒度的手部动作。从表 2 可看出，本文方法的行为识别准确度为 60.625%，如果仅使用粗粒度的全局行为信息，其识别率稍有降低，为 56.875%，其识别性能和传统人工特征提取方法具有可比性。从实验数据也可看出，粗细粒度信息的融合能有效提高识别准确度。然而，手部细粒度信息的添加对识别准确度的贡献并不大，可能是因为左手关节点处于变化之中，以左手关节点为中心截取的视频只能反映手部细节信息，丢失了重要的运动轨迹信息。值得注意的是，如果仅对第 11～16 个行为（即玩游戏、躺到沙发上、行走、弹吉他、站起和坐下）进行识别，则识别准确率达 98%，这是因为第11～16 个行为间具有较大的差异，而数据集中的其他行为之间的差异则非常细微，如读书、写字、用笔记本电脑几个行为仅是手部动作有细微差别。实验结果说明，使用深度学习方法能够有效进行行为识别，尤其是当各行为动作差别较大时，识别率会得到显著提高。

表 2　本文方法与人工特征提取方法识别性能比较

Tab. 2　Performance comparison between artificial feature extraction method and proposed method

算法	识别率/%
LOP features[5]	42.5
Joint Position features[5]	68
Dynamic Temporal Warping[19]	54
2CNN2F	56.875
2CNN2F_J	60.625

3.3　网络深度对识别的影响

关于如何构造深度神经网络到目前为止仍没有规律可循，现有的网络均是基于研究者的经验和实验探索。本文通过构建含 3 层 CNN 和 4 层 CNN 的神经网络，即 3CNN2F_8 和 4CNN2F（如表 3 所示），探讨了网络深度对识别效果的影响。网络参数如表 1 所示。为了保证网络不过渡拟合，本实验使用 24×8×128×128 的视频序列作为神经网络的输入，即将规范化后的 192×128×128 视频，以 8 为步长，拆分成 24 个8×128×128 的视频段，同时输入到具有 24 个并行结构的神经网络，此处只考虑了粗粒度信息，没有融合细粒度信息。由表 2 可知，使用 3CNN2F_8 网络时的识别率为 52.5%，而使用 4CNN2F 的识别率为 58.75%。由于实验数据限制，本文难以提供更深或更浅深度网络的实验结果，现有结果可能意味着网络深度的增加对提高行为识别率有一定的影响，但若要增加网络深度，必须提供更多的训练样本以防止过度拟合。

表 3　不同网络中的参数配置及识别率

Tab. 3　Recognition accuracies in different networks with different parameters

实验网络	网络输入	L_{Seg}	L_{Stride}	识别率/%
2CNN2F	12×16×32×32	16	16	56.875
2CNN2F_J	24×16×32×32	16	16	60.625
3CNN2F_8	24×8×128×128	8	8	52.5
3CNN2F_4	47×8×128×128	8	4	56.875
4CNN2F	24×8×128×128	8	8	58.75

3.4　拆分步长对识别的影响

为了检验拆分步长对识别效果的影响，本文针对 3CNN2F 构建了两个不同输入的网络：3CNN2F_8 和 3CNN2F_4（网络参数如表 1，表 3 所示）。为简化处理，本次实验也只考虑粗粒度信息作为输入，因此 3CNN2F_8 的输入为 24×8×128×128 的视频序列，而 3CNN2F_4 的输入的大小为 47×8×128×128，即将规范化后的 192×128×128 视频，以步长为 4，拆分成 47 个 8×128×128 的视频段，拆分后，相邻两个视频段间有 4 帧的重复。实验结果如表 3 所示。由表 3 可知，步长为 8 时，识别准确率为 52.5%，而步长为 4 时，识别准确率为 56.875%。识别率得到有效提高，主要是因为一方面步长越小，拆分的视频段越多，深度网络需要的并行分支也越多，在横向上变的更宽，网络参数越多，网络的泛化能力越好；另一方面，步长的减小和拆分视频段的增加，同时也增加了训练数据，使网络训练效果更好。

4　结　论

鉴于深度视频可以描述物体的几何结构，而

且对光线、颜色不敏感，本文以深度视频为研究对象，采用传统的二维CNN方法构建深度神经网络，对MSRDailyActivity3D数据集中的行为进行分类识别。实验及结果表明，本文提出的基于CNN的深度学习方法能够对以深度视频表示的人体行为进行有效识别，对MSRDailyActivity3D数据集中行为差异较大的躺下、行走、弹吉他、站起和坐下5个行为的平均识别准确率为98%，对整个数据集上所有行为的识别准确率为60.625%。接下来，本文还对如何提高深度学习的识别率进行了一定的探索。研究发现减小拆分视频段的步长，融合粗粒度和细粒度的视频信息，适当增加网络深度均能有效提高深度网络的识别率。未来的研究方向将主要集中在从不同粒度，不同信息源等方面进行信息融合以提高人体行为识别率。

参考文献：

[1] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *Piscataway*, NJ: *IEEE*, 2005: 886-893.

[2] TIAN Y L, CAO L L, LIU Z C, *et al.*. Hierarchical filtered motion for action recognition in crowded videos [J]. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2012, 42(3): 313-323.

[3] 张迪飞，张金锁，姚克明，等. 基于SVM分类的红外舰船目标识别[J]. 红外与激光工程，2016，45(1)：167-172.
ZHANG D F, ZHANG J S, YAO K M, *et al.*. Infrared ship-target recognition based on SVM classification [J]. *Infrared and Laser Engineering*, 2016, 45(1): 167-172. (inchinese)

[4] LI W, ZHANG Z, LIU Z. Action recognition based on a bag of 3D points [C]. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW), *Piscataway*, NJ: *IEEE*, 2010: 9-14.

[5] WANG J, LIU Z C, WU Y, *et al.*. Mining actionlet ensemble for action recognition with depth cameras [C]. 2012 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), *Piscataway*, NJ: *IEEE*., 2012: 1290-1297.

[6] XIA L, AGGARWAL J K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera [C]. 2013 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). *Piscataway*, NJ: *IEEE*, 2013: 2834-2841.

[7] OREIFEJ O, LIU Z. Hon4d: histogram of oriented 4D normals for activity recognition from depth sequences [C]. 2013 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). *Piscataway*, NJ: *IEEE*, 2013: 716-723.

[8] ZHANG C Y, TIAN Y L. Edge enhanced depth motion map for dynamic hand gesture recognition [C]. 2013 *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (CVPRW). *Piscataway*, NJ: *IEEE*, 2013: 500-505.

[9] YE M, ZHANG Q, WANG L, *et al.*. A survey on human motion analysis from depth data [J]. *Time-of-Flight and Depth Imaging, Sensors, Algorithms, and Applications*, *Springer*, 2013: 149-187.

[10] 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]. 2011 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). *Piscataway*, NJ: *IEEE*, 2011: 3361-3368.

[11] ZHANG N, PALURI M, RANZATO M, *et al.*. Panda: pose aligned networks for deep attribute modeling [C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). *Piscataway*, NJ: *IEEE*, 2014: 1637-1644.

[12] TOSHEV A, SZEGEDY C. Deeppose: human pose estimation via deep neural networks [C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). *Piscataway*, NJ: *IEEE*, 2014: 1653-1660.

[13] LIU P, HAN S, MENG Z, *et al.*. Facial expression recognition via a boosted deep belief network [C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). *Piscataway*, NJ: *IEEE*, 2014: 1805-1812.

[14] HE K, ZHANG X, REN S, *et al.*. Spatial pyramid pooling in deep convolutional networks for visual recognition [C]. *Computer Vision-ECCV 2014*, *Springer*, 2014: 346-361.

[15] LIN M, CHEN Q, YAN S. Network in network [J]. *Computer Science*, 2014.

[16] SZEGEDY C, LIU W, JIA Y Q, *et al.*. Going deeper with convolutions [C]. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*(CVPR),2015:1-9.

[17] 陈芬, 郑迪, 彭宗举,等. 基于模式复杂度的深度视频快速宏块模式选择算法[J]. 光学 精密工程, 2014, 22(8):2196-2204.
CHEN F, ZHENG D, PENG Z J, *et al.*. Depth video fast macroblock mode selection algorithm based on mode complexity [J]. *Opt. Precision Eng.*, 2014, 22(8):2196-2204.(inchinese)

[18] COLLOBERT R, KAVUKCUOGLU K, FARABET C. Torch7: A matlab-like environment for machine learning [R]. *BigLearn, NIPS Workshop*, 2011.

[19] MÜLLER M, RÖDER T. Motion templates for automatic classification and retrieval of motion capture data [C]. *Proceedings of the* 2006 *ACM SIGGRAPH, Eurographics Association*, 2006: 137-146.

作者简介:

刘　智(1977—),男,江西高安人,博士,副教授,2011 年于四川大学计算机科学与技术专业获得博士学位,主要从事深度学习、人体行为识别、图像处理、目标跟踪、信息融合研究。E-mail: liuzhi@cqut.edu.cn