DOI:10.19651/j.cnki.emt.1802315

基于骨架模型的人体行为分析

朱凌飞 万旺根
(1.上海大学 通信与信息工程学院 上海 200444; 2.上海大学 智慧城市研究院 上海 200444)

摘 要: 随着深度学习运用到图像领域,姿态估计、行为分析等算法的性能得到显著提升,希望在利用较好模型基础上进一步分析,在尽可能短的时间内得到更直观的结果。2016 年提出的沙漏堆网络对人体关节点进行多尺度、多阶段的训练,在 MPII 数据集上回归了 16 对关节点坐标,在单个 11 G 显存的 GPU 上的平均准确率为 87.6%;连接关节点构建人体骨架模型,然后根据骨架模型的加权角和倾斜角等几何特征,进一步推断人体的动作和行为状态,最后对人体行为进行分类和判断,包括站立、直坐、躺下等常见 7 类动作,平均准确率为 82%,优势在于有效降低计算量和处理时间。

关键词: 神经网络;姿态估计;行为分析;沙漏堆模型;几何特征
中图分类号: TP391;TN911.73 **文献标识码**: A **国家标准学科分类代码**: 520.2060

Human behavior analysis based on skeleton model

Zhu Lingfei Wan Wanggen
(1. School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China;
2. Institute of Smart City, Shanghai University, Shanghai 200444, China)

Abstract: With the application of deep learning in the field of image, the performance of algorithm such as pose estimation and behavior analysis has been significantly improved. We hope to further analyze based on better models and get more intuitive results in the shortest time. The Stacked Hourglass network proposed in 2016 carried out multi-scale and multi-stage training on human keypoints, and regressed 16 pairs of coordinates of keypoints on the MPII dataset with 87.6% average accuracy rate on single GPU of 11G video memory. These keypints were connected into a human body skeleton model. The motion and behavior of the human body are further inferred based on the geometric features of the skeleton model such as weighted angle and tilt angle. The result is classification on human behavior for seven common class actions including standing, sitting, lying and so on. The final average accuracy is 82% and the advantage is to effectively reduce the amount of calculation and processing time.

Keywords: neural network; pose estimation; behavior analysis; stacked hourglass model; geometric features

0 引 言

在视频监控和人机交互等计算机视觉研究中,对于人体行为识别与分析是目前热门的研究领域[1]。在对图像中人体行为分析的研究中,有些方法是用整幅图像直接进行分类,如基于空间金字塔法、随机森林法等构建的分类器;有些方法是利用对象与人之间的相互作用或是人体的姿态进行分类;还有些方法是利用整体和局部属性进行识别[2]。其中基于姿态的人体行为分析方法,首先需要在图像或视频序列中确定人体重要关节点的像素坐标,然后对人体行为进行分类。

人体姿态估计算法主要可分为基于整体的和基于部件的两大类[3]方法。基于整体的人体姿态估计算法一般通过待处理图像到部位或关节定位的非线性映射来实现,基于部件的人体姿态估计算法根据部位或关节间的外观和位置的关联建立人体模型,并通过优化由人体模型构造的能量函数来实现。早期的算法侧重于使用图像特征和复杂的结构化预测。可变形的组件模型(deformable component model,DPM)是一种有效的目标检测算法,建立人体模型和目标进行匹配是许多分类器、对象分割、人体姿态和行为分类的基础[4]。

随着深度学习的发展,由于深度学习模型具有很强的非线性变换能力,不仅大幅提高了图像识别的精度,同时也避免了需要消耗大量时间进行人工特征提取的工作,从大

收稿日期:2018-11-22

数据中学习具有更强表达和区分能力的特征,使得最终性能和运算效率大大提升[5]。

2014年提出的姿态识别从全局出发,对整个图像直接回归人体关节点坐标,对于每一个关节点,都把整个图像作为输入,使用卷积神经网络提取全局特征[3]。卷积姿态机器(convolutional pose machines,CPM)用各部件响应图来表达各部件之间的空间约束,把响应图和特征图一起作为数据在网络中传递,通过热力图回归人体关节点坐标[6]。

沙漏堆模型使用全卷积网络回归人体关节点坐标,思路明晰、网络简洁、训练时间减少[5]。网络采用多尺度分析的方法,捕捉并巩固了图像的多尺度信息,包含多个降采样和升采样的过程,结构对称,直观上像多个沙漏堆叠在一起。在训练时进行跨尺度自下向上、自上向下的误差传递和重复推断,2016年在世界范围内取得了非常好的结果[7]。

获得了人体关节点的信息之后,在进一步对行为对象进行特征提取和分析的过程中,有国内学者提出了行为对象特征和场景特征相结合的方法,更好地利用图像的所有信息进行人体行为分类,平均准确率为67.8%,比仅仅分析行为对象的特征提高了6%左右[8]。

本文采用人体骨架模型对人体行为进行分类。骨架作为目标物体图像一种紧致和拓扑结构的直观表形式,通常由分割后的二值图像中提取得到或者从灰度图像中直接提取得到。传统的骨架提取算法大多数针对目标的二值图像,保持原始图像的拓扑结构不变,具有单个像素宽度的属性,而且可以重新构造原始的目标。具体方法包括连续方法、精确方法和离散方法等。骨架的几何特征可以用面积、周长最长尺度、宽度、平均直径、外观比例、面积等价直径、形状因子和圆形性等测度来表示[9]。

本文在相关研究基础上,首先完成了沙漏堆网络模型的训练和部分参数的调优,在测试集上取得了87.6%的准确率。在此基础上,本文着重提出人体骨架模型的分析方法,计算了加权角、倾斜角等几何特征,充分利骨架的拓扑结构,进行了人体行为的分类的判断。本文的研究意义在于对深度神经网络模型测试结果的处理,大大降低了处理时间和复杂度,把模型输出的结果作为中间值,沿用人体行为分类的传统方法,取得了较好的结果。

1 关节点定位

1.1 沙漏堆网络

沙漏堆网络结构主要包含的模块如下:

1)输入层:对原始图像数据进行一定预处理,通常包括零均值化、归一化和白化等,大大降低了网络在训练时产生梯度消失等问题的影响。

2)卷积层:输入通道数设置为64或128,表示有64或128个不同的卷积核(过滤器、滑动窗口),对图像进行卷积,权值参数采用共享机制。卷积窗口大小如3×3或5×5等小尺寸,降低计算量,防止性能退化。这一层提取了图像初步特征。

3)批处理层:为了保证神经网络的正常训练——迭代和更新参数,需要保证样本数据的稳定性,在前向传递过程中,通过规范化数据分布。通过使用零均值化和方差归一化等方法,减小数据之间的差异性,避免因为部分离群样本导致参数进入冻结区域,提高训练的效率以及最终模型的性能。

4)残差模块:传统的卷积神经网络模型,都以层叠卷积层的方式提高网络深度来提高识别精度,但是网络过深时,误差反向传递会无法有效地把后面的梯度更新到前面的网络层,导致前面的层参数无法更新,即梯度消失或梯度弥散问题。残差网络(ResNet)包含了高维卷积路和跳跃路,在进一步提取图像高维特征的同时,通过跳跃连接把后面的梯度传递到前面的网络层中,从而保证了深度网络的正常训练,如式(1)所示。

$$y = F(x,\{W_i\}) + x \tag{1}$$

式中:x 表示输入;F 为残差函数,对中间函数进行参数学习。

5)池化层:介于连续的卷积层中间,用于压缩数据和降低参数数量,防止过拟合。对于一张图像,池化操作改变图像的大小,对于每个n×n窗口,得到一个数值作为输出值。窗口大小为2×2,即池化后长度和宽度都缩小为原始的一半,总像素缩小为原始的1/4。池化方法使用最大池化方式,即保留每个窗口中的最大值作为输出。

6)沙漏网络:不同于一些其他方法通过使用独立的网络在不同分辨率下处理图像,沙漏网络的特点在于采用特定的机制来有效地处理多尺度的特征,并在整个网络中把各部分结果合并。分支结构很好地保留了每个分辨率的空间信息,其中最小的尺度可以达到的4×4像素。假设原始尺度为256×256,那么处理的所有尺度可以用式(2)来表示。

$$R = \{p \times p \mid p = 2^n, n = 2,3,4,\cdots,\log_2 256\} \tag{2}$$

式中:p 表示图像的高度或长度。

沙漏网络中,卷积和池化层把图像特征降低到很小的分辨率,池化时关闭网络分支,每次降低一次尺度,池化前经过多次卷积提取特征。当多次池化到达最低的分辨率之后,网络开始重新恢复原始尺度,采用最近邻方法逐步向上采样,把不同分辨率下的特征进行组合。沙漏网络的拓扑结构是对称的,对于向下采样的每一层,都有相应的向上采样层,就如同两头大中间小的沙漏形状。

7)输出层:当尺度向上采样恢复原始分辨率后,再进行连续卷积产生最终的网络预测,输出为一组预测热图,可以预测每一个关节点出现在每个像素点的概率。

以上这些基本模块构成了一个沙漏结构,整个网络是通过将多个沙漏结构端到端的连接在一起,其中每一级的输出作为后一级的输入,从而把网络结构进一步加深[10]。一级沙漏网络结构如图1所示,实际网络由这个结构进行

拓展。在通过每个沙漏结构后生成一次预测，把该尺度和其他尺度下的特征结合起来处理。后一级的沙漏对高级特征进行再次处理，进一步评估更高层次的空间关系。网络采用多次迭代和中继监督的思想，具有重复自下向上的、自上向下的推理能力，能够对整个图像的初始估计和初始特征进行重新评估，降低损失。

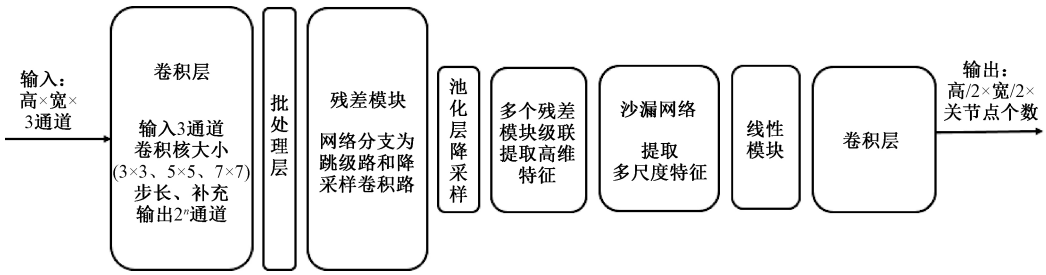

图 1　沙漏堆网络结构(一阶)

1.2　训练和测试过程

中继监督是在每个阶段的输出上都计算损失，这样可以保证底层参数正常更新。如果直接对整个网络进行梯度下降，输出层的误差经过多层反向传播会大幅减小，即很可能发生梯度消失现象。训练时，初始化参数可以随机设定，在适当的位置计算出初始的预测集。

除了在最下层开始向上采样时，大多数高阶特性只在较低的分辨率下呈现。如果在网络上采样之后进行中继监督，那么这些特性就无法在更大的全局环境中进行重新评估[11]。为了使预测更精确，仅局部范围内评估是不够的，关键在于分析每个关节点与其他关节点之间的关系以及对图像整体背景的理解来联合预测。还有一种方法是在池化步骤之前进行中继监督，但是这样对于给定像素点的特征，是通过处理相对局部接受域得到的，因此忽略了关键的全局信息。

在每个沙漏模块中结合了局部和全局的信息，并且网络早期就生成初步预测结果，要求它对图像有较高的理解，尽管只有部分是在整个网络中进行解析的，推断过程的后续阶段需要对这些特征进行更深入的重新学习。

对于一个如姿态估计的结构化问题，输出本质上是许多不同特征相互作用的结果，并且要使这些特征结合起来形成对场景的统一理解，因此在尺度之间来回切换的方法是关键。

梯度下降算法是神经网络模型训练最常用的优化算法。目标函数 $J(\theta)$ 关于参数 θ 的梯度为目标函数上升最快的方向。对于降低损失函数这种最小化优化问题，只需要将参数沿着梯度相反的方向前进一个步长，就可以实现目标函数的下降如式(3)所示。

$$\theta \leftarrow \theta - \eta \nabla_{\theta} J(\theta) \tag{3}$$

式中：θ 表示初始参数；η 表示步长或称为学习速率；$J(\theta)$ 表示目标函数；$\nabla_{\theta} J(\theta)$ 表示当前参数的梯度。

AdaGrad 是学习速率自适应的梯度下降算法。在训练迭代过程，其学习速率是逐渐衰减的，经常更新的参数其学习速率衰减更快。本文使用的 RMSprop 是对 AdaGrad 中学习速率可能过快问题的改进，借鉴冲量梯度下降 Momentum 的思想，引入一个超参数，在积累梯度平方项进行衰减，可以用式(4)、式(5)表示。

$$s \leftarrow \gamma \cdot s + (1-\gamma) \cdot \nabla_{\theta} J(\theta) \odot \nabla_{\theta} J(\theta) \tag{4}$$

$$\theta \leftarrow \theta - \frac{\eta}{\sqrt{s+\varepsilon}} \odot \nabla_{\theta} J(\theta) \tag{5}$$

式中：γ 为衰减参数，范围为 0 到 1，一般取值 0.9，有效减少了出现的爆炸情况[12]。S 表示学习率自适应更新参数；$\nabla_{\theta} J(\theta) \odot \nabla_{\theta} J(\theta)$ 为梯度的平方，仅仅对距离时间较近的梯度进行积累，有效避免学习速率很快下降的问题，学习速率一般设置为 0.001 或以下。

训练数据集 MPII 是通过一个既定分类系统对 YouTube 视频中每天人类活动进行收集的一个数据集，包含了大约 25 000 张图像，注释的人体关节点在 40 000 以上。数据集涵盖了 410 个人类活动，每张图像都有一个活动标签，并提供了前续和后续未加注释的帧，其中测试集还包含了身体遮挡、3D 躯干和头部朝向等更丰富的注释。

在处理的图像中，通常包括多个可见的人，如果没有图模型或其他后续处理步骤，图像必须向网络传输所有必需的信息，以确定注释的是其中的哪个人，即对这个人进行关节点的定位。这里使用一个英文字母来表示我们确定哪一个人，比如“D”表示从左往右的第 4 个人。

2　算法设计

通过神经网络在训练集上的训练，得到了更新完的模型参数，并在测试集上进行测试和展示，可以直观地查看关节点的位置。把这些关节点按照一定规律连接起来，就可以得到近似的人体骨架模型，如图 2 所示，由 16 个人体关节点组成 15 条骨骼线段。

本文对于骨架模型主要使用到的几何特征为加权角、切斜度和长度。

目标物体的区域重心是一种基于目标区域的全局描述方法，重心位置可以根据研究区域内的目标物体所有点计算得到，记为 G，如式(6)所示。

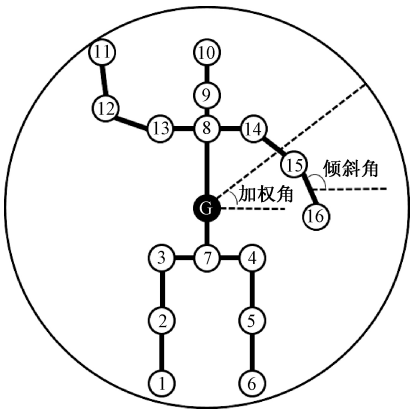

图 2 人体骨架模型几何特征

$$G(\bar{x},\bar{y})=\left(\frac{1}{A}\sum_{(x_i,y_i)\in R}x,\frac{1}{A}\sum_{(x_i,y_i)\in R}y\right) \tag{6}$$

式中:G 为重心;A 为关节点个数;(x_i,y_i) 为第 i 个关节点的坐标。

以 G 为中心作圆,要求覆盖整个目标物体,所以半径一般取目标物体上的点到重心的最远距离,可表示为式(7)。

$$r=MAX\left(\sqrt{(x_i-\bar{x})^2+(y_i-\bar{y})^2}\right) \tag{7}$$

式中:$(\bar{x},\bar{y})$ 表示中心坐标;r 表示半径。

对于骨架上的一条线段 L_i,会和某一条半径产生一个交点或者重合,这样就可以得到线段 L_i 对于重心的加权角,为 0°~360°中的某个值,进而确定线段的位置[13]。

骨架上的一条线段的倾斜度,可以用倾斜角的正切值来衡量。骨骼的倾斜度直接反应出了人体某个部位所处的状态,而且对于一些关键的躯干,如脊柱、大腿等,它们的切斜度直接决定了人体的行为状态,可以很轻易地以此为根据判断站立、平躺等行为。根据图 2 所示骨骼倾斜度结合加权角,实现对人体行为的分析。分析流程如图 3 所示。

其中对骨骼的分析是从主要到次要来进行选择的,如先分析脊柱,确定整个人行为的大体状态,再分析大腿、小腿骨骼,确定下肢的运动状态,最后对手部和头部进行分析。每次分析都把结果保存在状态表示数组中,最后通过大量的实验测试,找到合适的阈值来判断最终的人体行为。

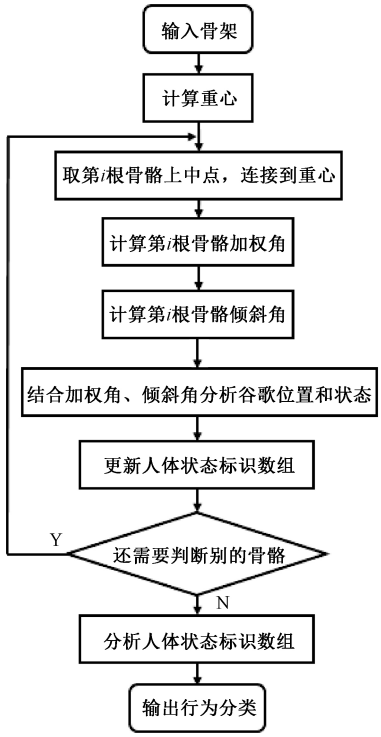

图 3 基于骨架模型几何特征的行为分析流程

3 实验结果与分析

3.1 实验 1:沙漏网络模型训练和测试

PyTorch 是 Python 语言中使用 GPU 和 CPU 优化的深度学习张量库,通过调用这个库的方法来建立沙漏网络模型并进行训练。CPU 为 Intel Xeon® CPU * 5650@2.67 @GHz * 24; GPU 为 Nvidia Geforce GTX 1080 Ti,显存 11 GB;内存为 24 GB;操作系统为 Ubuntu16.04;GPU 驱动程序版本为 Nvidia v384;CUDA 版本为 8.0;cuDNN 版本为 5.1.2; Python 环境为 Anaconda3; 优化器为 RMSprop;批大小为 6;初始学习率为 1.2×10^{-4};总训练次数为 120 次,其中第 60 次时学习率衰减为 1/10 即 1.2×10^{-5},第 90 次时学习率再衰减为 1/10 即 1.2×10^{-6};训练时间为 3 天多。统计结果如表 1 所示,测试数据集的平均准确率为 87.6%,部分结果如图 4 所示。

表 1 关节点定位测试准确率 (%)

方法	头部	肩部	肘部	腕部	臀部	膝盖	脚踝	平均
Tompson et al., NIPS′14	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira et al., CVPR′16	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al., CVPR′15	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu&Ramanan., CVPR′16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	72.4
Lifshitz et al., ECCV′16	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
本文方法	96.6	95.4	89.4	83.0	88.4	82.2	77.8	87.6

图 4　关节点定位测试结果

从实验结果来看，对于测试集中的大部分图像，在遇到遮挡、复杂背景、光照、姿态多样性和尺度角度不同等各种情况下，预测值的结果都比较接近真实值，其中人体上半部分的关节点定位性能要明显高于下半部分[14]。通过研究发现，性能较差的情况，基本出现在身体被图像截断这种情况。当图像中的人体不完整时，人体的关节点很可能出现在图像的外部，这时想在图像内部定位出关节点，显然误差会较大。

3.2　实验 2：骨架模型分析和行为分类

通过对骨架模型中骨骼的加权角、倾斜角和长度等几何特征的分析，把人体行为分为最常见的 7 类基本动作[15]，包括“站立”、“直坐”、“跪地”、“俯身”、“后仰”、“躺下”和“趴下”。这些类别的划分是在 MPII 的一部分测试集上，选择了具有代表性的 100 张图像，通过多人主观判断后，选择平均值进行标注。测试结果如图 5 所示，平均准确率为 82%，部分结果如图 6 所示。

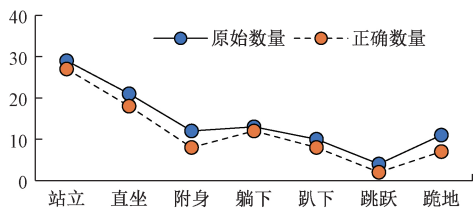

图 5　行为分析实验结果统计图(100 张图像)

从实验结果来看，在分析“俯身”和“后仰”，“躺下”和“趴下”这两对动作时，它们骨架的整体结构是非常相似，所以只能通过腿部和手部结构分析出人体的朝向。

此外，还可以发现，相对于“站”或“躺”等简单动作，“坐”、“跪”等复杂动作准确率较差，这是由于人体这类复杂动作可能会导致部分关节点位置出现错位和重叠的情况，毕竟图像只能包含人体的两维信息，丢失了深度信息。所以在这种情况下，即使关节点的位置仍然相对精确，但是相对于一般情况，骨架模型是变形的，它的一部分几何特征在用作分类依据时，可靠性会大大降低。

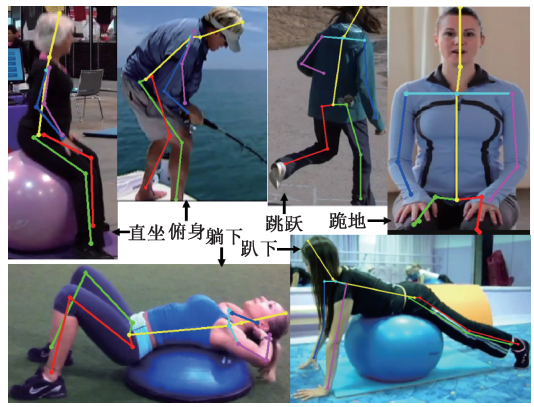

图 6　基于骨架模型几何特征的行为分析结果

相对于把行为对象特征和场景特征相结合的方法，本文算法只聚焦于人体行为特征，但是由神经网络提取的几何特征在性能上远优于传统方法，实验结果在性能上取得较大的突破，基本实现了人体行为分析的目标，证明了对模型输出结果分析这一思路的合理性。

4　结　　论

本文提出了一种基于骨架模型的人体行为分析方法。首先利用全卷积神经网络沙漏堆网络获得人体的关节点位置，然后构建人体骨架模型并分析模型的几何特征，最后对人体行为进行分类和判断[16]。通过实验可知，深度神经网络模型的训练需要计算能力较强的设备，网络的参数需要在多次训练的过程中评估最好的一组数值，相对于传统方法，关节点的定位在测试集上性能优越。在此基础上进行骨架几何模型分析，可以有效降低计算量和处理时间，能够把结果进一步转化为对于人体行为的分析，这实际上是对于神经网络训练结果的分析和利用[17]。当然，分析的过程还可以继续细化，实验的样本数量也有待提高。在后续的实验过程中，一方面需要适量改进网络的模型结构，另一方面需要利用更多的几何特征进行度量，从而判别出更多的行为类别。

参考文献

[1]　RAMAKRISHNA V, MUNOZ D, HEBERT M, et al. Pose machines: Articulated pose estimation via inference machines[C]. Computer Vision-ECCV, 2014.

[2]　JAIN A, TOMPSON J, LECUN Y, et al. MoDeep: A deep learning framework using motion features for human pose estimation[C]. Computer Vision-ACCV, 2014:302-315.

[3]　赵勇，巨永锋. 基于改进卷积神经网络的人体姿态估计[J]. 测控技术，2018，37(6)：9-14.

[4]　TOSHEV A, SZEGEDY C. Deeppose: Human pose estimation via deep neural networks [C]. CVPR,

IEEE, 2014:1653-1660.

[5] JAIN M, JÉGOU H, GROS P. Asymmetric hamming embedding: taking the best of our bits for large scale image search[C]. ACM International Conference on Multimedia, 2011:1441-1444.

[6] 马淼,李贻斌.基于多级动态模型的2维人体姿态估计[J].机器人,2016,38(5):578-587.

[7] 苏延超,艾海舟,劳世竑,图像和视频中基于部件检测器的人体姿态估计[J].电子与信息学报,2011,33(6):1413-1419.

[8] JÉGOU H, CHUM O. Negative evidences and cooccurences in image retrieval: The benefit of PCA and whitening[C]. European Conference on Computer Vision, 2012:774-787.

[9] EIGEN D, PUHRSCH C, FERGUS R. Depth map prediction from a single image using a multi-scale deep network [J]. Advances in Neural Information Processing Systems,2014:2366-2374.

[10] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.

[11] 史东承,冯占君.视频中人体行为分析[J],吉林大学学报(信息科学版),2014,32(5): 521-527.

[12] 黄永鑫.基于视觉的运动人体行为分析技术研究[J].黑龙江科技信息,2010(27):24.

[13] 李春霞,杨克俭,李波.人体骨架模型的建立及IK问题的一种解决方式[J].武汉理工大学学报(交通科学与工程版),2003,27(6):815-818.

[14] TOMPSON J, JAIN A, LECUN Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[J]. Eprint Arxiv, 2014:1799-1807.

[15] JOHNSON S, EVERINGHAM M. Learning effective human pose estimation from inaccurate annotation[C]. Computer Vision and Pattern Recognition (CVPR), 2011:1465-1472.

[16] 唐超,王文剑,李伟,等.基于多学习器协同训练模型的人体行为识别方法[J].软件学报,2015,26(11):2939-2950.

[17] 冉宪宇,刘凯,李光,等.自适应骨骼中心的人体行为识别算法[J].中国图象图形学报,2018,23(4):519-525.

作者简介

朱凌飞,硕士,主要研究方向为图像视频处理。

E-mail:zhulingfei2015@126.com

万旺根,博士生导师,主要研究方向为计算机图形学、信号处理和数据挖掘。

E-mail:wanwg@staff.shu.edu.cn