带特征监控的高维信息编解码端到端无标记人体姿态估计网络

沈 栎,陈 莹
(江南大学轻工过程先进控制教育部重点实验室,江苏无锡 214000)

摘 要: 针对点云空间三维信息非结构化和旋转易变性对预测结果的影响,提出一种带特征监控的三维信息编解码卷积神经网络,该网络可实现三维空间下单目深度图的端对端无标记人体姿态估计. 所设计的网络由特征监控编解码组件串联而成,该组件第一部分使用三维卷积模块以类似沙漏结构的形式组合设计,实现对特征图的编码和解码;第二部分以不同参数残差块并联,实现对特征图的监控融合,第一部分与第二部分首尾连接构成组件. 特征监控编解码组件能根据数据集大小,通过串联的方式搭建不同深度的网络,同时根据数据分辨率,设置组件参数,实现由粗到精的特征学习,最终获得最佳网络. 通过 ITOP 数据库的实验表明,该网络实现了空间三维信息的端到端深度学习,显著提高了系统性能并具有更高的精度.

关键词: 计算机视觉; 深度图; 人体姿态估计; 深度学习; 三维数据卷积网络
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2020)08-1528-10
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2020.08.010

Feature Monitored High-Dimension Endecoder Net for End to End Markless Human Pose Estimation

SHEN Li, CHEN Ying
(*Key Laboratory of Advanced Control Light Process, Jiangnan University, Wuxi, Jiangsu* 214000, *China*)

Abstract: Aiming at the impact of unstructured and rotational variability of three-dimensional information in point cloud on prediction results, a feature-supervised three-dimensional information encoding and decoding convolution deep learning network is proposed. The network is composed of feature monitoring coding and decoding modules in series. In the first part of the module, a three-dimensional convolution module is used in the form of hourglass structure to realize the coding and decoding of the feature map. In the second part, the residual blocks of different parameters are connected in parallel to realize the monitoring and fusion of feature maps. Feature monitored coding and decoding modules can build networks with different depths in series according to the size of data sets. At the same time, according to the data resolution, modules parameters can be set to realize feature learning from rough to fine, and ultimately obtain the best network. The experiment of ITOP database shows that the network achieves the end-to-end deep learning of three-dimensional information, significantly improves the system performance and has higher precision accuracy.

Key words: computer vision; depth image; pose estimation; deep learning; 3D-CNN

1 引言

精确的三维人体姿态估计是行为识别,人机交互接口领域的基础研究课题,在增强现实、动画制作等领域具有广泛的应用前景. 通过单目深度图的进行无穿戴,无标记人体姿态估计,是该领域一个重要的分支,其主要的特点是不使用具有标记作用的穿戴设备,仅利用单目深度图进行人体关节点的估计. 由于人体的姿态多样性,变化的不确定性以及信息的非结构化等特性,基于单目深度图的人体姿态估计存在着由自遮挡、扭曲、畸变等因素造成的估计误差,且计算复杂度较高.

与基于 RGB 彩色信息的深度学习[1]不同,基于单

收稿日期:2019-05-28;修回日期:2019-11-19;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61573168)

目深度图的人体姿态估计方法具有空间信息,方法可分为两类,分别是基于2D深度图特征的方法及基于3D点云特征的方法. Shtton[2]等人利用深度空间信息作为辅助设计深度图特征,实时的估计出单目情况下深度图中的人体关节点. Shafaei[3]等合成多张深度图,通过岭回归的方法估计关节点,避免了单目情况下的点云残缺,提高了关节点估计的精度. Haque[4]等建立的深度学习网络,能从二维深度图中,估计出三维空间的关节点位置.

将深度图变换为点云后再基于3D点云特征进行姿态估计工作,避免了维度转换产生的透视畸变及非线性映射[5]. Moon[6]等提出了一种可以进行三维信息卷积的模块,在此基础上,提出了一个基于HourGlass网络[7]改进的三维点云深度学习网络V2V-PoseNet[5],将其应用在手势估计领域,有效的回归出了三维空间关节点,由于V2V-PoseNet只进行一次编解码操作,导致网络的深度固定,针对不同训练数量的可控性较低,整体网络结构僵化.

为了学习到姿态估计更精细的特征,提高预测结果,本文提出了一种端到端三维信息深度学习组件,该组件称为特征监控编解码器 Monitored-Endecoder(ME). ME可实现对三维数据进行卷积,避免了因维度畸变产生的误差;此外,ME通过对特征图的特征进行监控融合,加强了特征图的非线性,提取到了更精细的特征,提高了预测精度. 并且,ME对于不同大小的数据集,能以串联的方式建立不同深度的网络,并能根据数据分辨率设置监控参数,实现由粗到精的特征学习. 本文通过将监控参数逐渐增大的ME组件进行串联,建立了一个端对端的三维信息卷积神经网络,称为"带特征监控的高维编解码网络 Feature Supervised High-dimension Endecoder Net(FeSHEN)",该网络可用于在三维空间对单目深度图中无标记人体的关节点进行估计,经过ITOP数据集验证,相比之前的方法具有更高的预测精度.

2　框架设计

本节阐述了算法的总体框架和点云提取的方法,并对点云提取的作用进行了分析.

2.1　算法总体框架

如图1所示本文算法的总体框架. 输入为一幅2D单目深度图像,上分支应用 Deeprior ++[8]网络输出参考质心与真实质心的偏差值;下分支将二维深度图映射形成点云,通过简单深度阈值方法计算参考质心;综合上下分支数据,利用偏差值修正参考质心得到预测质心并完成标准化,之后使用以质心为中心的固定大小立方体 ***Cubic*** 提取人体区域,***Cubic*** 称为体素集,该集合所在空间称为体素空间.

将 ***Cubic*** 输入本文所设计的 FeSHEN 网络,通过端到端的深度学习,得到每一个关节点在体素空间下的特征热图 HeatMap,其取值最大的点,即为体素空间下预测关节点的位置. 最终将特征热图映射到世界坐标,得出世界坐标系下各个关节点位置. 本文的工作,主要集中在 ***Cubic*** 之后的部分,设计了一种新的三维卷积模块 ME,基于该模块提出了一种全新的网络结构 FeSHEN,实现了基于三维信息端到端深度学习方法的单目深度图的无标记人体姿态估计.

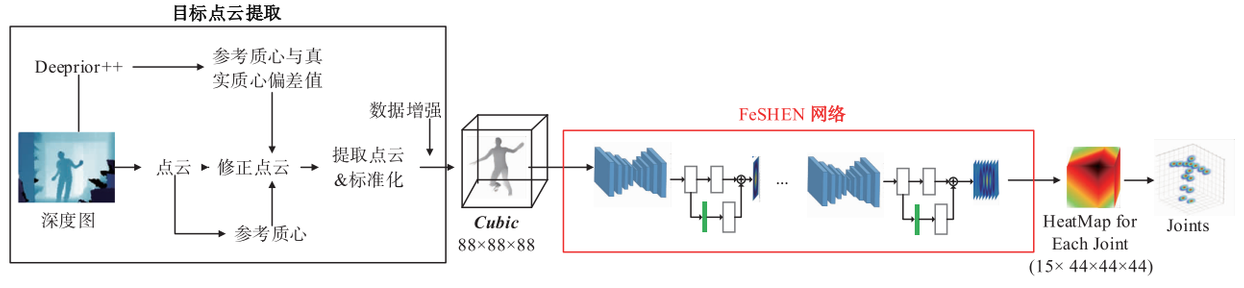


图1　算法总体框架

2.2　目标点云提取

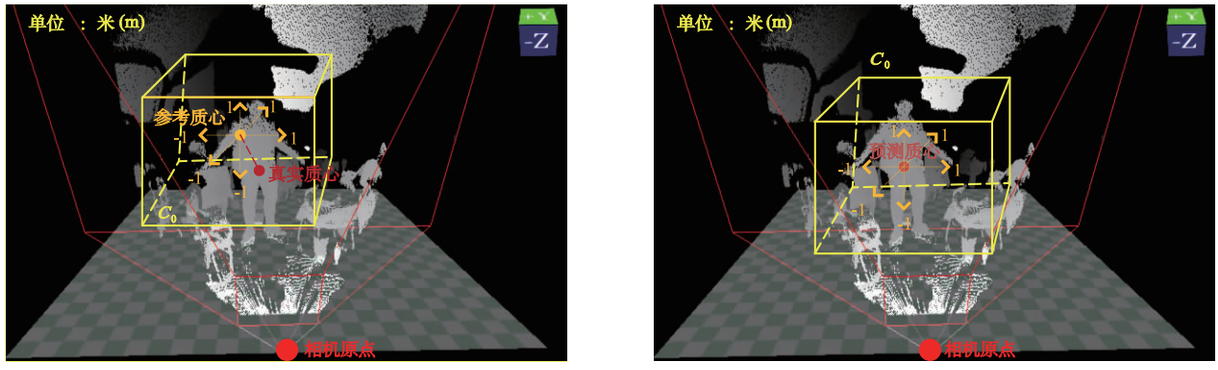
如图2所示,红色圆点为相机坐标系原点,点云各点坐标以原点确认,该坐标系下的单位为米(m). 由于使用数据的采集设备为 Asus Xtion Pro 相机,根据相机的工作范围限定,x,y 取值范围[-1,1],z 取值范围[0.8,3.5].

定义立方体 ***Cubic*** 为 $\boldsymbol{C}_0$,原点为参考质心. $\boldsymbol{C}_0$的 x,y,z 轴取值均为[-1,1],如图2所示. 通过将 $\boldsymbol{C}_0$的原点与人体的真实质心重合,人体被完全包裹起来,达到人体检测的目的.

常用的质心坐标的计算方法如质心真实值参考[9],及简单深度阈值提取目标区域质心[10],在实际应用中无法得到真实值,且在杂乱场景下,当两个对象距离过近时,因为对所有输入数据应用相同的阈值,无法保证获取到正确的质心位置. 因此,$\boldsymbol{C}_0$可能会只包含目标对象的某些部分,导致提取的人体不完整,如图2(a)

所示.

为了解决这个问题,按照 Oberweger[8] 等提出的卷积神经网络 Deeprior ++ 方法训练了一个简单的卷积网络,来获得简单深度阈值[10]计算的质心和真实质心间的偏差值 *xyzOffset*. Deeprior ++ 网络的核心思想是利用启发式的训练,将主成分分析(PCA)计算的三维先验信息直接集成到卷积神经网络中. Deeprior ++ 首先利用参考质心[10]作为网络输入,输出为参考质心与真实质心的三维偏移量 *xyzOffset*. 经过三维偏移校正后的 $\boldsymbol{C}_0$ 如图 2(*b*)所示.



(*a*) 偏移量作用前 (*b*) 偏移量作用后

图2 对质心的修正

2.3 点云空间与体素空间的映射

点云空间到体素空间具体转换过程:

设输入点云集 $\boldsymbol{P}=[\boldsymbol{x}_p,\boldsymbol{y}_p,\boldsymbol{z}_p]$,根据简单阈值计算的参考质心为 C_{Ref},参考质心与真实质心的偏差值为 *xyzOffset*,

$$\begin{cases}\boldsymbol{P}_{\mathrm{in}}=\boldsymbol{P}-C_{\mathrm{Pre}}\\ C_{\mathrm{Pre}}=C_{\mathrm{Ref}}+xyzOffset\end{cases}\tag{1}$$

其中,C_{Pre}为预测质心的位置. 根据式(1)获得修正后的源数据 $\boldsymbol{P}_{\mathrm{in}}$. 通过减去质心坐标,将点云质心移动到了相机原点处,将体素空间 $\boldsymbol{C}_0$的原点设为点云质心,此时可计算点云在体素空间下的坐标,即实现了空间的转换.

然后标准化:

$$\boldsymbol{P}_{\mathrm{out}}=\frac{\boldsymbol{P}_{\mathrm{in}}-(C_{0,Min})}{C_{0,Max}-C_{0,Min}}\times 88\tag{2}$$

其中 $C_{0,Min}=-1$,$C_{0,Max}=1$,是 $\boldsymbol{C}_0$的最小值与最大值,该步骤将相机坐标系下的点云坐标转换到体素坐标系 $\boldsymbol{C}_0$下. 归一化后乘以 88 进行标准化,将 $\boldsymbol{C}_0$大小调整为 $88\times88\times88$.

根据 $\boldsymbol{P}_{\mathrm{out}}=[\boldsymbol{x}_{\mathrm{out}},\boldsymbol{y}_{\mathrm{out}},\boldsymbol{z}_{\mathrm{out}}]$,将 $\boldsymbol{C}_0$中对应坐标点的位置数值设为 1,无坐标的位置数值设为 0,最终生成网络输入 ***Cubic***.

3 带监控编解码器与带特征监督网络

本节先讲解网络核心 ME 组件的具体构造,分析了 ME 组件在卷积过程中对特征图的作用机制,以及监控参数的调整对学习到的特征的细化程度的影响. 之后介绍 FeSHEN 网络的整体架构,搭建细节,以及反向传播损失函数细节.

3.1 带监控编解码器组件

ME 组件是一种体素到体素的预测模型,为 CNN-3D 结构,将空间中 *Z* 轴作为额外空间轴,因此 Kernel 为三维卷积核.

3.1.1 组件构造

在 Hour-Glass 模型[9]基础上,受到 Pavlakos[11]中解构过程监控融合的启发,设计了一个用于三维体素估计的网络组件 ME,具体结构如图 3 所示.

该组件由池化块,残差块,反卷积块和监督块组成,为了简化图形,没有将 *Z* 轴信息在图中表示出来. 每一个模块下面的两行分别表示该模块输入输出,数字代表该模块处理数据的空间大小和特征的通道数,模块内立方体的尺寸代表特征图的大小,立方体的厚度代表通道的大小. 图 3 中最左边的虚线框部分是编码器,池化块用来减小特征图的尺寸大小,残差块用来增加特征通道数量,每一个卷积核与图像卷积后的处理结果存放在一个通道 channel 内,因此通道的数量与卷积核的数量相同. 模型遵从先验的超参数设定,在二维情况下,主流模型的通道数主要为 32 及 64,针对较复杂的三维数据,将该模型最大通道数设置为 128.

中间虚线框部分是解码器,反卷积块在增加特征图的空间大小的同时,用更少的卷积核作用特征图,减少通道数,实现压缩和解码. 在编码过程中,通过较小的步长来压缩特征,通过增加通道数来扩大特征学习数量,使网络特征更丰富,更容易下降到密集性关键点位置,从而定位关节点.

最右边的虚线框部分代表监控器,主要由不同参数的残差块组成,通道参数记为 Out 输出参数,和 Moni 监控参数. 该监控器有两个分支,均由两个残差块组成. 上分支用于对特征图进行通道变换,残差块的通道数为参数 Out;下分支的第一个残差块,为中间监控模

块(Intermediate Monitor Block),该模块的输入通道数为通道参数 Out,输出通道数为监控参数 Moni,通过该残差块,将特征图进行压缩,提取非线性特征;下分支的第二个残差块,输入通道数为监控参数 Moni,输出通道数为输出参数 Out,主要用于对压缩后的监督特征进行通道扩展,进而与原输出特征进行融合. 监督参数即监控块输入输出的卷积核的数量,根据点云单位空间内数量进行设定,单位空间内点云的点越多,则监督参数可以适当增大. 但是因为监控块主要功能是学习点云的大略特征,因此该参数设置不宜过大. 本文中该参数根据网络的深度按文献[2,4,8,16]的数值依次增大.

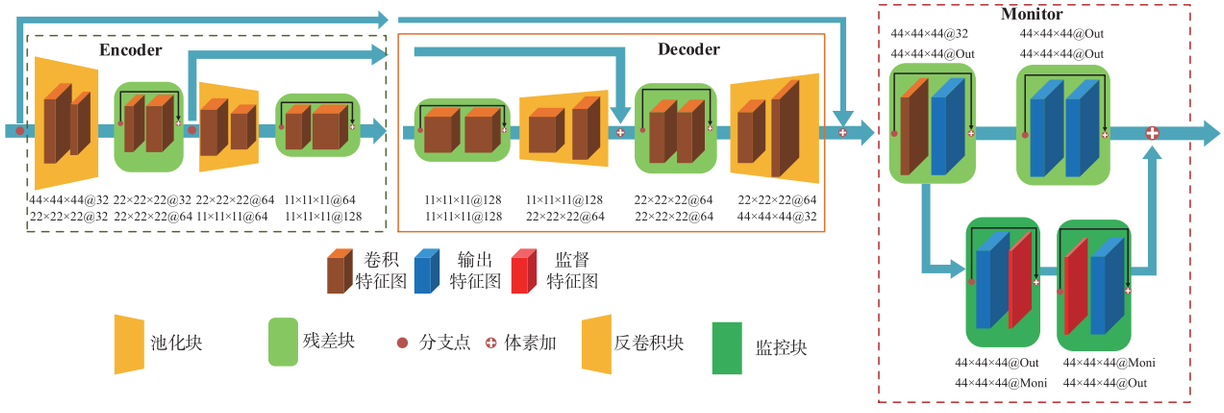


图3　ME组件

3.1.2　组件分析

图 3 所示,ME 组件经历了上采样(Decoder)的过程,根据 Odena[12] 团队的研究,二维图像在反卷积之后,像素点会呈现出方格状的棋盘形式,这种情况被称为棋盘效应(Checkerboard Artifacts),其原因是步长与卷积核不能整除造成重叠特征反馈. 在三维情况下,会形成重叠区域的三次累加,本质上比二维情况更容易影响预测结果.

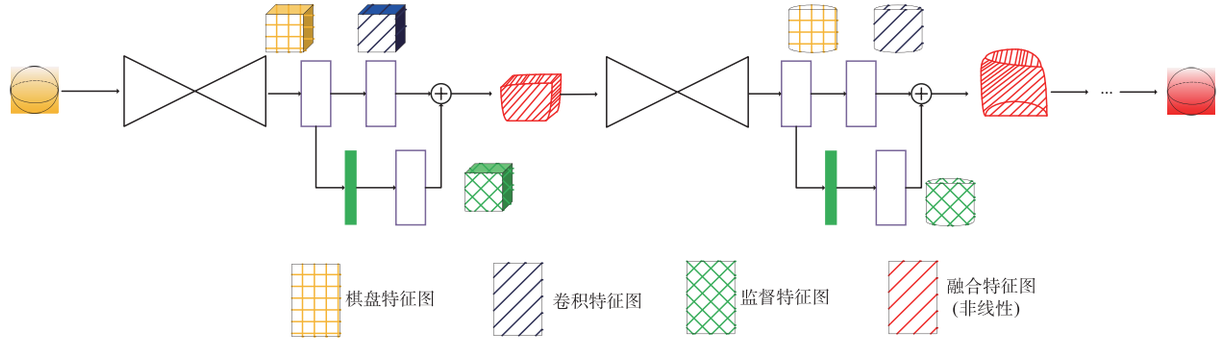


图4　Monitor作用过程

如图 4 所示为监控模块的作用示意图. 输入为一个球体时,经过上采样,产生了棋盘效应,生成了黄色方块特征图,图中方格线条代表棋盘状. 上分支对这个方块进行卷积,生成的蓝色卷积特征图,其中斜线代表特征图发生了一定的非线性变化. 下分支对其进行监督压缩和扩展,生成绿色监督特征图,其中交叉线表示该特征图具有更多的非线性. 将绿色监督特征图和蓝色卷积特征图融合,输入红色融合特征图,与黄色方块相比增加了非线性.

为获取更为精确的 3D 特征图,在 Endecoder 模块之后添加 Monitor 部分,用于消除 3D 棋盘效应的影响,输出非线性更强的特征图. 此外,通过串联 ME 组件,实现对网络深度的调整,对于不同数据集大小可以做出更改. 组件的监控参数,代表监控模块的输入和输出的通道变化,根据文献[11]的解释,用输出通道从小到大的模块,可实现从大略特征到主要特征,再到细化特征的学习. 因此,在串联过程中,针对三维数据的分辨率不同,使用不同模块输出通道数,实现从粗到精的特征学习,选取不同细化程度的数据特征,不断进行非线性操作,逼近原始特征.

3.2　带特征监督的高维信息编解码网络

本节介绍了构建网络的基本单位,网络的具体结

构,损失函数的选取和参数更新的算法.

3.2.1 网络设计

构建网络的基本单位为 3D 卷积基本块,共 4 种. 第 1 种基本块由体素卷积、体素批归一化[13](Batch Normalize)层、激活函数(ReLU)组成,称为卷积块,主要进行三维信息的卷积计算;第 2 种基本块,由对文献[14]中用于二维图的 2D 残差模块进行 3D 体素拓展得到的,称为残差块,主要用于对三维特征图的通道数进行调整;第 3 种基本块用于对 3D 体素进行下采样,其原理和 2D 池化是一样的,称为池化块,主要用于三维降低特征图的尺寸大小;第 4 种基本块用于对 3D 体素进行上采样,由体素上采样(Upsampling)、体素 BN 层、ReLU 组成,称为反卷积块. 在卷积块和反卷积块中使用批归一化层和激活函数有助于简化学习过程,加快下降速度. 在网络设计过程中,残差块的 Kernel 大小为 $3\times3\times3$,卷积块和反卷积块的 Kernel 大小为 $2\times2\times2$,步长均为 2.

基于上述模块,本文针对三维体素信息,设计了一个带监督的高维信息编解码网络(FeSHEN),该网络的主要部分由四个 ME 组件使用不同的监控参数串联组成. 通过对 ME 组件进行串联,加深网络层数,利用不同的监控参数,实现对不同细化程度的特征进行提取. 具体的网络结构如图 5 所示.

图 5 中模块下的数字,代表该模块处理特征图的尺寸与通道数,其中红色数字代表监控参数 Moni,ME 组件的监控参数依次为[2,4,8,16],蓝色数字代表输出参数 Out,ME 组件的输出参数依次为[8,16,32,64].

网络的输出为在体素坐标系下,输出的对应第 n 个关节点的特征热图 $\boldsymbol{He}^n$,其最大值将出现在第 n 个关节点的体素坐标处.

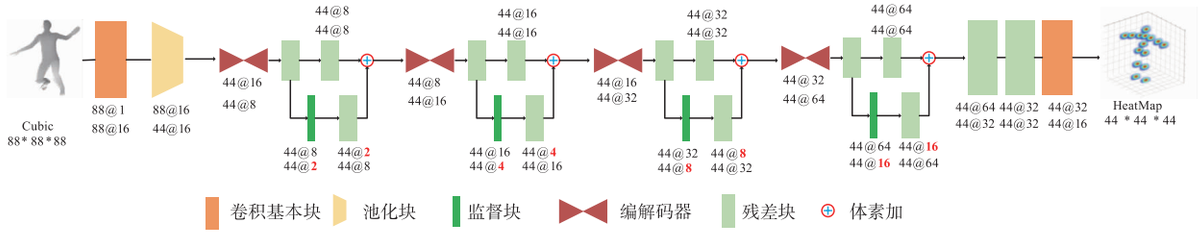


图5 FeSHEN网络

在测试阶段,为获取目标关节点坐标,令

$$\boldsymbol{J}^n_{\text{in}} = \underset{i,j,k\in[1,\cdots,44]}{\text{argmax}} \boldsymbol{He}^n(i,j,k) \tag{3}$$

$\boldsymbol{J}^n_{\text{in}}$是对应第 n 个关节点的特征热图 $\boldsymbol{He}^n$ 中最大值点对应的体素坐标

根据2.3 节的映射关系,反向即可求取目标关节点的坐标,具体根据以下公式:

$$\boldsymbol{J}^n_{\text{out}} = \boldsymbol{J}^n_{\text{in}}/44\times(Max-Min)+Min+C_{\text{Pre}} \tag{4}$$

3.2.2 损失函数

采用预测值与真实值的均方误差作为损失函数 L,具体如下:

$$L = \sum_{n=1}^{N}\sum_{i,j,k}\| H^*_n(i,j,k) - H_n(i,j,k)\|^2 \tag{5}$$

式(5)中,N 代表关键点的标记,H^*_n 和 H_n 分别是第 n 个关节点的真实值和预测值的热图表示.

为了考虑到对每个点的预测可能性,真实值热图表示 H^*_n 通过在真实关节点处建立高斯极值点得到,具体公式如下:

$$H^*_n(i,j,k) = \exp\left(-\frac{(i-i_n)^2+(j-j_n)^2+(k-k_n)^2}{2\sigma^2}\right) \tag{6}$$

其中,$(i,j,k)\in[1,\cdots,44]$是预置热图的坐标,i_n,j_n,k_n 是第 n 个关节点在点云空间的真实值,$\sigma=1.7$ 是高斯极值点的标准差.

网络的优化过程采用 RMSProp 方法[15],相比于 AdaGrad 的历史梯度,RMSProp 增加了一个衰减系数来控制历史信息的获取多少,被证明是有效且实用的深度学习网络优化算法.

4 实验结果及分析

本节先给出了验证数据集和具体实施细节,然后对实验结果进行了可视化分析,接着比较了不同数量组件串联的预测精度,再同相关方法做了更细致的比较,最后给出同之前主要方法比较的数据.

4.1 数据集与实施细节

本文使用的是 Haque 等公开的 ITOP 数据集[4]. 该数据集对应正面视角包含有 40k 个训练图像和 10k 个测试图像,所使用的采集设备为 Asus Xtion Pro 相机. 对应每一幅图,给出了图中 15 个人体关节点的 3D 空间坐标值.

具体网络实施采用 Torch7[16]框架搭建,在 NVIDIA 1080 TI GPU 上进行训练测试. 训练过程权重采用 $\mu=0.001$ 的零均值高斯分布进行初始化,使用均方误差(MSE)作为损失函数,用 RMSProp[15]方法进

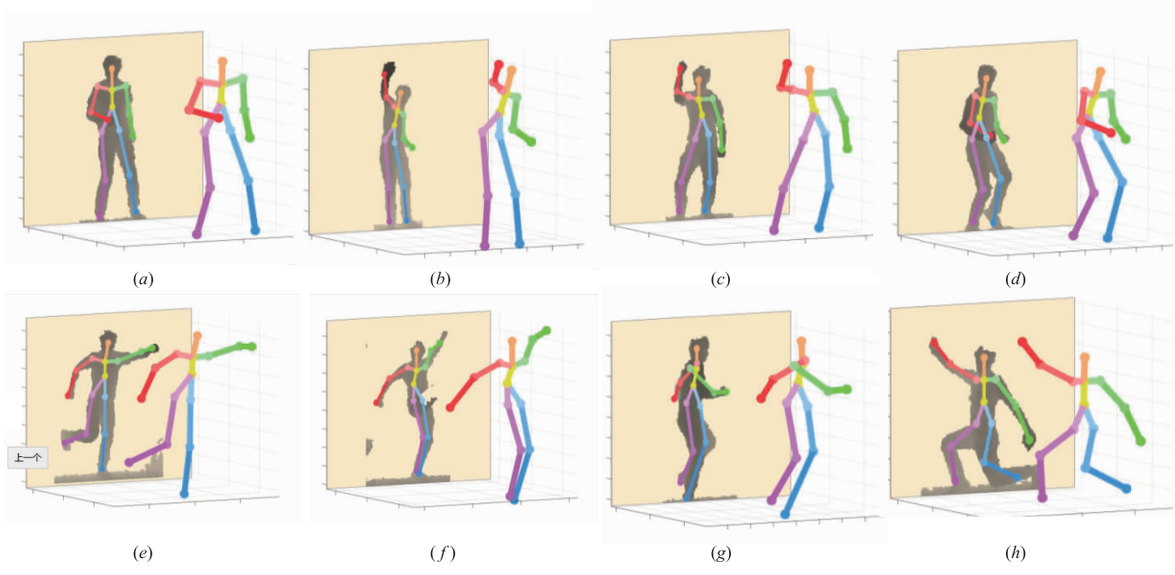
行权值更新,其 mini-batch 的尺寸设置为 8,学习率设为 2.5×10^{-4}. 为了得到更为精确的数据,训练 10 个模型,在预测时,综合不同模型的结果求均值来获得最终的预测值.

4.2　可视化分析

将预测结果转换到二维和三维空间与原图进行匹配,实现预测结果可视化.

4.2.1　复杂动作

如图 6 所示为对于无明显自遮挡的正面人体运动的预测结果,对于正常的人体正对相机情况,如(*a*)~(*d*),网络给出了准确的关键点预测. 对于复杂动作,如(*e*)~(*f*),在有足够训练量的基础上,FeSHEN 网络也能输出相对正确的结果. 虽然在某些关节点处会存在些许误差,如(*f*)中膝部.

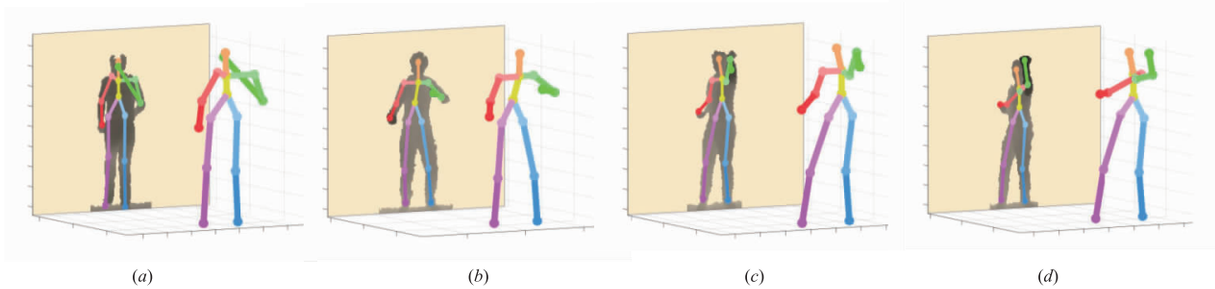


(*a*)　(*b*)　(*c*)　(*d*)　(*e*)　(*f*)　(*g*)　(*h*)

图6　复杂动作预测

4.2.2　细微动作

如图 7 所示为一个连贯的人体动作. 当手举起,由于自遮挡,产生关节点抖动,如图(*a*). 当手部举到一半,手部与身体分离,关节点预测回归正确,见图(*b*). 手部继续上升,又产生人体自遮挡,导致点云歧义,手部关节点预测产生误差,见图(*c*). 当手部完全举起,消除歧义,关节点回归正确,此时右肩被自遮挡,从图像检测域消失,出现微小误差,但接近于真实值,证明 FeSHEN 具有一定鲁棒性,见图(*d*).



(*a*)　(*b*)　(*c*)　(*d*)

图7　细微动作分析

4.2.3　困难动作

图 8 示出了一个连续的预测序列. 当人体处于站立时,预测效果较好,对于人体逐渐下蹲的动作,由于动作复杂性逐渐升高,当下蹲到一定的程度,预测结果失真,导致错误,如图(*d*)、(*e*). 但本文网络针对较大动作的变化仍能示出相关节点,见图(*b*)、(*c*). 由于训练样本中,缺少此类复杂动作的样本,因此最终的预测结果,出现了错误. 但网络本身具有一定的鲁棒性,对于畸变,仍然能预测出大部分的关键点位置. 就算是如图(*d*)、(*e*)中的未知样本,对于仍具有相关特征的关键

点,如足部,手部,仍然能做出正确的预测.

4.3 数据验证

本节先介绍了串联不同 ME 组件验证的实验结果,之后在 ITOP 的人体正面数据集上,比较了将 FeSHEN 网络和 V2VPoseNet 网络的预测精度,最后给出了相比之前方法 FeSHEN 网络的验证精度.

(a) (b) (c) (d) (e)

图8 困难动作预测

4.3.1 ME 组件的串联特性

本文在 ITOP 数据库人体正面数据集上验证了不同数量 ME 组件串联的预测结果. 为了增加数据的可信度,训练了三个模型,求取各个关键点输出的均值作为预测值,将预测值与真实值之间的欧氏距离作为误差,表 1 列出了各个关节点在测试样本集中误差的均值和标准差.

由表中可见,对于本文所用数据集,当模块串联数量为四个时,预测效果为最佳. 随着网络层数的增加,学习到更多的数据特征,消除了过拟合影响. 当串联数量过多,细化程度超过数据分辨率基础单位,出现欠拟合,精度又下降.

表 1 不同数量模块串联预测结果

误差(cm)	均值与标准差				
串联数量	1	2	3	4	5
头	3.30 ±0.224	3.45 ±0.205	**3.21** ±0.212	3.39 ±0.202	3.39 ±0.206
颈	5.01 ±0.215	3.79 ±0.222	3.84 ±0.208	**3.71** ±0.205	4.24 ±0.207
肩	6.54 ±0.371	5.30 ±0.376	5.37 ±0.440	**4.94** ±0.391	5.60 ±0.373
肘	8.17 ±0.817	**7.96** ±0.683	8.36 ±0.828	8.24 ±0.800	8.22 ±0.815
手	12.41 ±1.387	12.37 ±1.319	11.91 ±1.290	**11.43** ±1.269	12.10 ±1.381
腰	4.97 ±0.231	4.76 ±0.245	5.75 ±0.212	**4.43** ±0.209	5.03 ±0.390
臀	8.49 ±0.445	6.83 ±0.382	7.20 ±0.414	**6.77** ±0.421	7.26 ±0.413
膝	6.10 ±0.628	6.26 ±0.613	6.13 ±0.621	**6.01** ±0.572	6.16 ±0.624
足	7.75 ±1.062	**7.17** ±0.995	7.73 ±1.073	7.68 ±1.008	7.28 ±0.890
Mean	6.97 ±0.598	6.43 ±0.621	6.61 ±0.588	**6.29** ±0.564	6.59 ±0.589

4.3.2 相关方法比较

针对 ITOP 人体正面数据集,比较了 FeSHEN 网络与 V2V-PoseNet 网络预测结果与真实值之间欧氏距离的均值和标准差,单位为厘米(cm),具体见表 2. 可见 FeSHEN 的均值误差较小,标准差也较小.

可见相比于 V2V-PoseNet 网络,本文提出的网络整体表现都优于相关网络. 图 9 中对比了两个网络模型的预测结果,图 9(Ⅰ)中所示某一人体运动序列,当人体的运动导致足部出了深度相机检测范围,V2V-PoseNet 做出了错误的估计,使足部关键点回到了初始位置,FeSHEN 给出了更接近真实值的预测值. 图 9(Ⅱ)中,(a)(b)两组序列,对于足部关节点,V2V-PoseNet 将两个点预测为一个点,而 FeSHEN 则将其细化的分开了,体现了 FeSHEN 具有更强的非线性预测效果. 而图 9(Ⅱ)中(c)(d)(e)三组序列,显示了 FeSHEN 具有更强的鲁棒性,针对人体复杂的运动,FeSHEN 能更准确的对变化的深度图信息做出反应,调整预测值,结果更接近于真实值.

表 2 与 V2V-PoseNet 比较

误差(cm)	均值与标准差	
比较网络	FeSHEN(Ours)	V2V-PoseNet
头	**3.39** ±0.202	3.60 ±0.203
颈	**3.71** ±0.205	4.33 ±0.213
肩	**4.94** ±0.391	5.50 ±0.431
肘	8.24 ±0.800	**8.08** ±0.781
手	**11.43** ±1.269	11.63 ±1.398
腰	**4.43** ±0.209	4.70 ±0.314
臀	6.77 ±0.421	**6.61** ±0.386
膝	**6.01** ±0.572	6.02 ±0.639
足	**7.68** ±1.008	8.16 ±1.022
Mean	**6.29** ±0.564	6.51 ±0.599

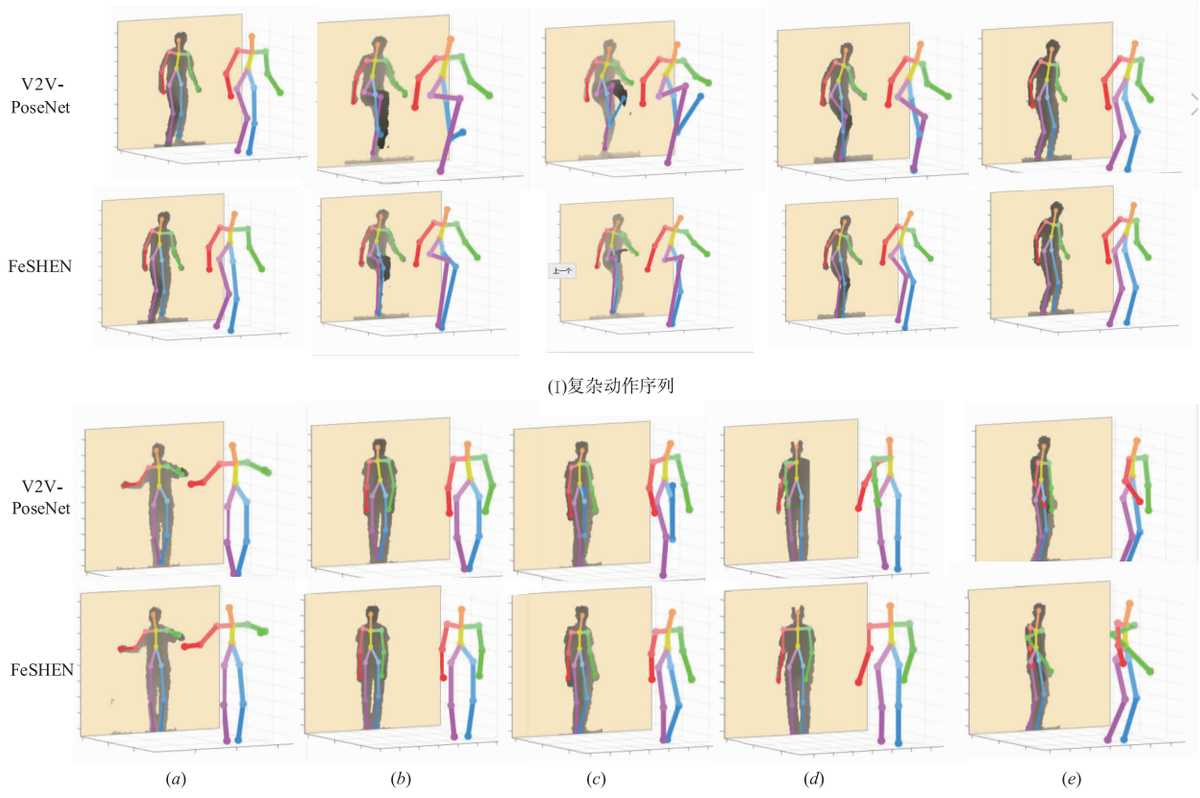


图9　与V2V-PoseNet的对比

4.3.3　同其他方法比较

对于每一个预测关节点,计算 3D 空间下,预测值落在真实值周围 10cm 范围内的概率. 该指标称为平均精度准确率(mAP)[3,17].

本文在 ITOP 3D 姿态估计数据集[4]上同先前的工作进行了比较. 其中包括基于随机森林的方法(RF)[2],RTW[17],IEF[18],基于视点不变特征的方法(VI)[4],REN-9×6×6[19],使用卷积存储机制的方法 CMB[20],基于锚点的三维估计方法 A2J[21]等. mAP 比较结果如表 3 所示,其中部分方法的数值记录来自于文献[5,10,19]. 表 3 中可见,本文的模型关注网络的可延展性和适应性,预测精度优于目前现有的方法,表明该网络可以应用于单目深度图的无标记人体姿态估计并具有更高的精度.

表 3　同其他方法的比较结果

	mAP								
比较方法(年份)	RF (2013)	RTW (2015)	IEF (2016)	VI (2016)	REN-9×6×6 (2017)	CMB (2018)	V2V-PoseNet (2018)	A2J (2019)	FeSHEN (Ours)
头	63.8	97.8	96.2	98.1	98.7	97.7	98.29	98.54	**98.96**
颈	86.4	95.8	85.2	97.5	**99.4**	98.5	99.07	99.02	99.14
肩	83.3	94.1	77.2	96.5	96.1	75.9	**97.18**	96.23	96.61
腕	73.2	77.9	45.4	73.3	74.7	62.7	80.42	78.92	**84.14**
手	51.3	70.5	30.9	68.7	55.5	**84.4**	67.26	68.35	71.56
腰	65.0	93.8	84.7	85.6	98.7	96.0	98.73	98.52	**99.06**
臀	50.8	80.3	83.5	72.0	91.8	87.9	**93.23**	90.85	92.77
膝	65.7	68.8	81.8	69.0	89.0	84.4	**91.80**	90.75	91.67
足	61.3	68.4	80.9	60.8	81.1	83.8	87.60	86.91	**89.75**
Mean	65.8	80.5	71.0	77.4	84.9	83.3	88.74	88.0	**91.52**

5 结语

本文提出了一种新的用于对单目深度图进行无标记人体姿态估计的端对端深度学习网络——有监督高维信息编解码网络(FeSHEN).该网络可以对三维信息进行深度学习回归,实现了更高的信息维度卷积计算,并能应用于端到端情况.相比于现有的方法,能更高精度的预测出单目深度图所对应三维空间下人体关节点的坐标.另外,本文设计了一种用于处理三维空间信息的三维卷积组件——带监控的编解码器组件(ME),该组件可以对三维空间信息进行卷积,并能够根据数据集的大小和数据库图片的分辨率,以串联的方式搭建不同深度的网络,适应不同大小的数据集,并用监控参数调整学习特征的细化程度,最终得到最优的网络结构.

参考文献

[1] 罗会兰,童康,孔繁胜.基于深度学习的视频中人体动作识别进展综述[J].电子学报,2019,47(5):1162-1173.
LUO Huilan,TONG Kang,KONG Fansheng. The progress of human action recognition in videos based on deep learning:a review[J]. Acta Electronica Sinica,2019,47(5):1162-1173.(in Chinese)

[2] Shotton J,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.

[3] Shafaei A,Little J J. Real-time human motion capture with multiple depth cameras[A]. Computer and Robot Vision(CRV)[C]. Victoria,Canada,2016. 24-31.

[4] Haque A,Peng B,Luo Z,et al. Towards viewpoint invariant 3d human pose estimation[A]. European Conference on Computer Vision(ECCV)[C]. Amsterdam,Netherlands,2016. 160-177.

[5] Moon G,Yong Chang J,Mu Lee K. V2V-posenet:Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)[C]. Utah,USA,2018. 5079-508.

[6] Moon G,Chang J Y,Suh Y,et al. Holistic planimetric prediction to local volumetric prediction for 3d human pose estimation[J]. arXiv preprint arXiv:1706.04758,2017.

[7] Newell A,Yang K,Deng J. Stacked hourglass networks for human pose estimation[A]. European Conference on Computer Vision(ECCV)[C]. Amsterdam,2016. 483-499.

[8] Oberweger M,Lepetit V. Deepprior ++:Improving fast and accurate 3d hand pose estimation[A]. Proceedings of the IEEE International Conference on Computer Vision(ICCV)[C]. Venice,Italy,2017. 585-594.

[9] Oberweger M,Wohlhart P,Lepetit V. Training a feedback loop for hand pose estimation[A]. IEEE International Conference on Computer Vision(ICCV)[C]. Santiago,Chile,2015. 3316-3324.

[10] Guo H,Wang G,Chen X,et al. Region ensemble network:improving convolutional network for hand pose estimation[A]. IEEE International Conference on Image Processing(ICIP)[C]. Beijing,China,2017. 4512-4516.

[11] Pavlakos G,Zhou X,Derpanis K G,et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[A]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)[C]. Hawaii,USA,2017. 1263-1272.

[12] Odena A,Dumoulin V,Olah C. Deconvolution and checkerboard artifacts[J]. Distill,2016,1(10):e3.

[13] Ioffe S,Szegedy C. Batch normalization:accelerating deep network training by reducing internal covariate shift[A]. International Conference on International Conference on Machine Learning(ICML)[C]. 2015. 448-456.

[14] He K,Zhang X,Ren S,et al. Deep residual learning for image recognition[A]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)[C]. Las Vegas,USA,2016. 770-778.

[15] Collobert R,Kavukcuoglu K C,Farabetin Biglearn. Torch7:A matlab-like environment for machine learning[R]. Neural Information Processing Systems Workshop,number EPFL-CONF-192376. 2011.

[16] Tieleman T,Hinton G. Lecture 6. 5-rmsprop:Divide the gradient by a running average of its recent magnitude[J]. COURSERA:Neural Networks for Machine Learning,2012:4(2):26-31.

[17] Jung H Y,Lee S,Yong S H,et al. Random tree walk toward instantaneous 3D human pose estimation[A]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)[C]. IEEE,2015. 2467-2474.

[18] Carreira J,Agrawal P,Fragkiadaki K,et al. Human pose estimation with iterative error feedback[A]. IEEE Conference on Computer Vision and Pattern Recognition(CVPR)[C]. IEEE,2016. 4733-4742.

[19] Guo H,Wang G,Chen X,et al. Towards good practices for deep 3d hand pose estimation[J]. arXiv preprint arXiv:1707.07248,2017.

[20] Wang K,Lin L,Ren C,et al. Convolutional memory blocks for depth data representation learning[A]. International Joint Conferences on Artificial Intelligence Organization(IJCAI)[C]. Stockholm,Sweden,2018. 2790-2797.

[21] Xiong F,Zhang B,Xiao Y,et al. A2J:Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image[J]. arXiv preprint arXiv:1908.09999,2019.

作者简介

沈　栎　男,1993 年 6 月出生于贵州都匀,江南大学硕士研究生,目前主要研究方向为三维图像处理和人体姿态估计.
E-mail:6161918015@vip.jiangnan.edu.cn

陈　莹　女,1976 年生于浙江丽水,江南大学教授,博士生导师,目前主要研究方向为图像处理、信息融合、模式识别.
E-mail:chenying@jiangnan.edu.cn