控制与决策
Control and Decision

基于改进DenseNet网络的人体姿态估计

石跃祥， 许湘麒

引用本文:
石跃祥, 许湘麒. 基于改进DenseNet网络的人体姿态估计[J]. 控制与决策, 2021, 36(5): 1206−1212.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2019.1218

基于改进DenseNet网络的人体姿态估计

石跃祥[1,2†], 许湘麒[1]

(1. 湘潭大学 信息工程学院，湖南 湘潭 411105；2. LED照明驱动与控制应用工程技术研究中心，贵州 铜仁 554300)

摘 要: 针对图像中由于人数不确定对处理速度的影响,以及不同人体或人体自身部位的相对大小不同等尺度因素影响导致通用的关键点检测方法的检测效果不佳等问题,提出一种改进的稠密卷积网络(DenseNet)结构用于人体姿态估计. 该网络结构为单阶段的端对端的网络结构,利用深度卷积神经网络进行特征提取,在卷积网络末端通过特定的尺度转换结构得到6种不同尺度的特征图,使得网络能同时使用不同层次的特征进行多尺度关键点检测,可以有效提高检测精度. 所提出方法采用自底向上的方式,使得网络进行多人姿态估计任务的处理速度得到保证. 实验表明,所提出方法相比几种主流方法在多人关键点检测的平均精度上提升了1个百分点,为平衡姿态估计的速度与精度提供了一种新方法.

关键词: 人体姿态估计；关键点检测；DenseNet；多尺度特征；尺度转换；深度学习

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2019.1218 **开放科学(资源服务)标识码(OSID):**

引用格式: 石跃祥,许湘麒. 基于改进DenseNet网络的人体姿态估计[J]. 控制与决策, 2021, 36(5): 1206-1212.

Improved DenseNet network for human pose estimation

SHI Yue-xiang[1,2†], *XU Xiang-qi*[1]

(1. College of Information Engineering，Xiangtan University，Xiangtan 411105，China；2. LED Lighting Research and Technology Center of Guizhou，Tongren 554300，China)

Abstract: In order to solve the problem that the impact of speed, and the poor detection performance of the common keypoints detection method caused by the uncertain number of people in the image and the relative size of different human bodies or body parts, an improved DenseNet network structure is proposed for human pose estimation. This network structure is a single-stage and end-to-end network, which uses deep convolutional neural networks for feature extraction. At the end of the convolutional network, it can get 6 different scales of feature maps by using a specific scale-transfer structure. Then the network can integrate different levels of features for multi-scale keypoints detection, which effectively improves the detection accuracy of keypoints. The bottom-up approach is adopted to ensure the processing speed of the multi-person pose estimation task. Experiments show that this method improves the mean average precision of multi-person keypoints detection by 1 % compared with other general methods. It provides a new method for balancing the speed and accuracy of attitude estimation.

Keywords: human pose estimation；keypoints detection；DenseNet；multi-scale feature；scale transformation；deep learning

0 引 言

在计算机视觉中,人体姿态估计是指基于图像信息对人体的各个关节和刚性部件进行准确检测和有效组合,其目的是获取人身体各个关键点的位置,得到正确位置后,对关键点进行正确的连接形成人体骨架信息,后续研究可以利用骨架信息对人的动作和行为进行分析. 在智能监控领域,利用人体姿态估计通过监控设备可对人的某些反常行为进行辅助监控和预警;可应用在体感游戏、虚拟显示等技术中,获取人的实时位置和姿态;可用于体育竞技中对人的动作分析;在智能人机交互方面,允许人可通过做出某种动作来与智能机器进行信息交互传递. 其应用前景广泛,具有很好的研究和应用价值,是计算机视觉领域备受关注的研究内容之一.

深度学习使得快速检测人的位置同时估计人的姿态成为一个重要且极具实用意义的任务,但是实现

收稿日期: 2019-08-28；**修回日期:** 2019-12-05.
基金项目: 国家自然科学基金项目(61602397, 61502407).
责任编委: 李少远.
†通讯作者. E-mail: shiyx@xtu.edu.cn.

人体姿态估计的过程中会出现一些无法避免的干扰因素. 人体姿态估计的主要难点如下:

1) 图像中人的数量事先是无法确定的,这会要求检测器对整个图像进行至少一次遍历,给检测器的检测速度带来挑战.

2) 不同的人所穿的衣服颜色不同,这会让人的部位外观存在多种可能的形式,且人的体型胖瘦和自身或其他物体的部分遮挡都会对检测器的检测精度带来挑战.

3) 图像中不同人离镜头的远近使得人与人之间相对图像的占比有所不同,不同的拍摄角度则会使得自身的各个部位相对图像的占比有所不同,这些均属于尺度因素的影响,常见的方法都是基于深度卷积神经网络实现的,其卷积核尺寸固定,即用于人体姿态任务的检测器尺寸是固定的,这对检测器检测不同尺度的特征有很大的难度从而影响人体姿态估计的结果,直接导致准确度下降.

2014年以前,人体姿态估计任务主要基于一些特定的结构模型,这些结构模型通过人体各个部件的约束关系建立[1-2],其灵活性和准确度均不高. 而随着深度学习相关理论的发展和完善,其基本取代了人工提取特征的工作,检测器的准确度也大大提高,所以目前人体姿态估计任务的主流是基于深度卷积神经网络的深度学习方法.

现阶段人体姿态估计任务按处理思路划分大致可分为两种:一种是自顶向下的方法,该方法先检测图像中的人,将人的位置框定,然后检测图像中已框定的人的各个关键点,得到的结果是包含所有关键点的一张热力图谱(Heatmap);另一种是自底向上的方法,这类方法直接检测图像中所包含的全部的人体关键点,然后将这些关键点确定到每个人的分组.

早期以单人姿态估计[3-5]为主,其中Toshev等[3]提出的DeepPose是最早应用深度卷积神经网络的方法,利用卷积网络进行特征提取,使精度得到了很大提升,之后大部分方法都基于卷积网络. Wei等[4]提出的卷积姿态机最具代表,很多方法均由该种方法改进而成,其主要贡献在于使用顺序化的卷积结构表达空间信息和纹理信息,对同一个卷积架构同时使用多个尺度处理输入的特征和响应,使得精度进一步提高,考虑了各个部位之间的远近距离关系. Newell等[5]提出的堆叠沙漏网络是设计一种与反卷积结合的沙漏结构,充分利用多尺度特征捕获人体各关节点的空间位置信息,堆叠多次对人体的关键点位置进行更为准确的推断,使得堆叠沙漏网络对关键点的检测精度明显优于卷积姿态机,但处理时间不尽如人意. Chou等[6]基于沙漏模型采用生成对抗方法训练网络,考虑了判别器的对抗损失使得模型处理效果好于单一的沙漏模型,推理时移除判别器,只用生成器进行推理保证了推理速度,但训练过程更为复杂,生成器和判别器训练时的同步性不易掌握.

随着深度卷积神经网络应用于单人姿态估计任务越来越成熟,很多研究人员开始在多人姿态估计任务中应用深度卷积神经网络. Insafutdinov等[7]使用残差卷积神经网络找出所有候选的关键点,对各个点进行标记并对关键点进行聚类,得到姿态估计结果,该方法在速度和准确度上相比前人方法有很大提升. ArtTrack方法[8]试图将姿态估计应用到视频跟踪里,通过简化稀疏身体部位关系图对单帧图像的处理速度进行提升,从而对视频帧进行快速处理. 其中Cao等[9]提出了一种多阶段两分支网络,由卷积网络进行特征提取,一个分支对所有关键点进行预测,另一个分支则预测关键点之间的部分亲和域(part affinity fields)以判断关键点之间的连接,使用偶图匹配的方法分配关键点到每个人,此方法是自底向上方法的代表,可以实现多人实时人体姿态估计,但并未解决遮挡与接触等问题,从而出现估计错误的情况,精度的上升空间很大.

针对人体姿态估计任务中通用方法精度与速度不匹配的问题,希望在提高精度的同时处理速度也得到提高,从而能够在要求高精度的实时任务中得以应用,有必要综合考虑人体姿态估计任务的速度与精度之间的均衡性以进行进一步的研究.

本文主要工作如下:提出一种基于卷积神经网络的单阶段、端对端的网络结构,并采用自底向上的方法快速进行多人人体姿态估计任务;对DenseNet卷积神经网络末端进行改进,设计特定的尺度转换结构,得到不同尺度的特征图,利用多尺度特征进行联合检测,提高了关键点的检测精度;与多种单人姿态估计方法和多人姿态估计方法进行速度和精度的实验对比,结果表明所提出的方法具有较大优势.

1 改进DenseNet

DenseNet[10]能够将前后层特征信息得到充分利用,从而获取更为丰富的特征图(feature map). 根据此特性,通过改进DenseNet设计一个单阶段、端对端的网络结构以进行人体姿态估计. 在进行卷积操作时,特征图越大所包含的图像细节越多,但拥有的高层次语义信息会越少,特征图越小具有的高层次语义信息越好,但分辨率会越差[11]. 将两者结合起来,可以在

得到很好细节的基础上,获得尽可能强的图像语义信息[12]. 在使用多个层次的特征进行联合检测时,能够使得尺度较小的目标也能有效地被检测到[13]. 因此,为了得到包含多个尺度的特征图信息,本文使用一种类似上采样(upsampling)的结构将特征图进行大小转换以得到多个尺度的特征图,然后使用不同尺度的特征进行关键点检测. 首先检测出所有关键点,并在检测的同时预测关键点之间连接成一个肢体部位的概率,然后通过偶图匹配的方法将关键点分配给每一个人,从而保证进行快速的人体姿态估计.

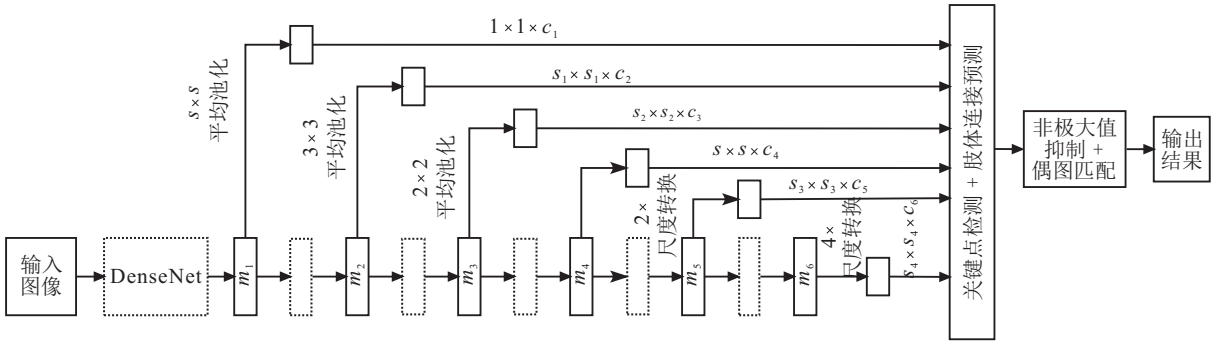

图1 网络结构

1.1 网络结构

利用特征金字塔解决多尺度问题,使得人体姿态估计准确度得到有效提升,但是处理速度较慢[5,14-15]. 为了获取多尺度特征信息,改进DenseNet网络并加入尺度转换结构得到不同大小的特征图组成特征金字塔,在特征金字塔上进行关键点检测,从而达到检测不同尺度特征的目标,提高网络对尺度不一的关键点的检测能力. 采用自底向上的处理思想,保证算法在进行多人姿态估计时的处理速度.

DenseNet可以通过前后特征的重复利用具有很强的特征提取能力,它有3种不同网络深度,分别为121层、169层和201层. 如图1所示,本文选取DenseNet最后一个密集连接模块的6个融合层进行改进,加入尺度转换操作以获取6个不同尺度的特征图. 利用这些特征图进行关键点的联合检测,网络在检测人体关键点位置的同时也会预测两个关键点连接形成肢体的概率,可在不使用多阶段网络处理[5,9,15]的情况下提高预测精度,同时快速地进行关键点检测.

对不同深度的DenseNet进行改进,在DenseNet的最后一个密集模块(第4个密集模块),利用尺度转换操作将特征图调整为6个不同尺寸的待检测特征图映射,如下所述:

1) 在第m_1个融合层输出$s \times s \times c_1$大小的特征图,经过一个$s \times s$的平均池化层得到$1 \times 1 \times c_1$大小的特征图;

2) 在第m_2个融合层输出$s \times s \times c_2$大小的特征图,经过一个3×3的平均池化层得到$s_1 \times s_1 \times c_2$大小的特征图;

3) 在第m_3个融合层输出$s \times s \times c_3$大小的特征图,经过一个2×2的平均池化层得到$s_2 \times s_2 \times c_3$大小的特征图;

4) 在第m_4个融合层输出$s \times s \times c_4$大小的特征图,该层不作处理,直接输出;

5) 在第m_5个融合层输出$s \times s \times c_5$大小的特征图,经过一个2倍尺度转换层得到$s_3 \times s_3 \times c_5$大小的特征图;

6) 在第m_6个融合层输出$s \times s \times c_6$大小的特征图,经过一个4倍尺度转换层得到$s_4 \times s_4 \times c_6$大小的特征图.

DenseNet-121中,$m_1, m_2, \ldots, m_6$对应最后一个密集模块中的第1、4、7、10、13、16个融合层,其所得的特征图大小均为$s \times s$,若输入图像尺寸为384×384,则得到输出为12×12,即s为12. $c_1, c_2, \ldots, c_6$对应第$m_1, m_2, \ldots, m_6$个融合层输出的特征图通道数,在此依次为544、640、736、832、232、64. $s_1, \ldots, s_4$表示在第m_2、m_3、m_5、m_6个融合层经过相关处理后最终输出的特征图尺寸,分别为4、6、24、48.

在DenseNet-169中,$m_1, m_2, \ldots, m_6$对应最后一个密集模块中的第5、10、15、20、25、32个融合层,若输入图像尺寸为384×384,则得到输出s为12. $c_1, c_2, \ldots, c_6$对应的值为800、960、1 120、1 280、360、104. $s_1, \ldots, s_4$的值与DenseNet-121相同.

在DenseNet-201中,$m_1, m_2, \ldots, m_6$与DenseNet-169相同,若输入图像尺寸为384×384,则得到输出s为12. $c_1, c_2, \ldots, c_6$对应的值为1 056、1 216、1 376、1 536、424、120. $s_1, \ldots, s_4$与DenseNet-121相同.

1.2 尺度转换结构

如图1所示,DenseNet最后一个密集模块的所有层的输出都具有相同的宽度和高度,只是通道数不同. 例如,输入图像尺寸为384×384时,最后一个密集模块所有的层输出都为12×12. DenseNet的网络结构能够将低层纹理信息通过密集连接汇集到融合层,所以融合层输出的特征图既有底层纹理信息又有高层语义信息,利用这一特性,设计了一个特殊的尺度转换结构(scale-transfer structure, STS),将其嵌入到DenseNet的融合层后直接改变输出的特征图尺寸,以获取多个尺度的特征图.

如图2所示,尺度转换结构会将输入的特征图在通道(channel)上按照r^2长度进行划分,即划分成C个,各通道长度为r^2的特征图,然后通过压缩输入特征图的通道数将每个$1 \times 1 \times r^2$小块展平转换为$r \times r \times 1$大小的块,最后将其堆叠成$rW \times rH \times C$的特征图,可以表示为

$$I_{\text{out}}(w,h,c) = I_{\text{in}}(\lfloor w/r \rfloor,\ \lfloor h/r \rfloor,\ r \times \text{mod}(h,r) + \text{mod}(w,r) + C \cdot r^2). \tag{1}$$

其中: I_{out}为转换后输出的特征图, I_{in}为输入的特征图. 与上采样不同的是,尺度转换结构避免了引入附加的参数量并有效降低了计算量,提高了检测器的检测速度. 在实验部分对此结构与上采样进行实验分析.

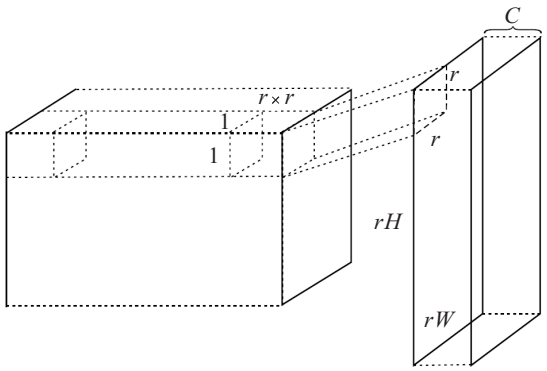

图 2　尺度转换结构

1.3 人体姿态估计

本文采用目标检测的思想[16-17]对关键点进行检测和连接,将输入的特征图分割为$M \times N$个网格,每个网格相当于一个特征图块,使用预测边框预测每一个网格,并给出一个区域建议(region proposal)的集合$\{D_k^i | k \in \gamma\}$. 其中: i表示第i个网格且$i \in \varphi = \{1,2,\ldots,M \times N\}$; γ表示要检测的目标且$\gamma = \{0,1,\ldots,K\}$,值为0表明目标预测为一个完整人,值$1 \sim K$作为关键点的标号. 如图3所示, D_k^i包含了预测边框的置信度和预测边框的坐标、宽度和高度.

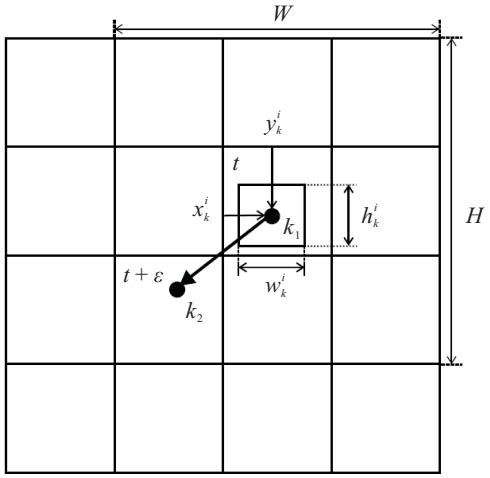

图 3　关键点连接预测示意图

预测边框的置信度为$\sigma_k^i = p(A|k,i)$,用以计算第i个网格检测到关键点k的概率,其中A为一个二值变量,如果关键点k的真实位置中心落入网格i中,则i中预测边框用来预测k. 用$\varphi_k^i = p(B|A,k,i)$计算第i个网格中能预测关键点k的边框与其对应真值区域的交并比(intersection over union, IoU),其中B为一个二值变量. 坐标(x_k^i, y_k^i)表示预测边框的中心相对于网格i边界的距离,以网格左上角为原点,对网格长度进行归一化, w_k^i和h_k^i表示网格i检测关键点k的预测边框的宽和高,分别以特征图像的宽和高进行归一化. 每个网格生成的区域建议D_k^i可表示为

$$D_k^i = \{\sigma_k^i, \varphi_k^i, x_k^i, y_k^i, w_k^i, h_k^i\}. \tag{2}$$

根据关键点的区域建议将检测到的关键点区域用矩形边框框定,框的大小与人的尺度成比例,并且始终用对应的关键点真值区域监督预测边框,有

$$R_k = \min\{(\sigma_k^i - \hat{\sigma}_k^i)^2 + (\varphi_k^i - \hat{\varphi}_k^i)^2\}. \tag{3}$$

在检测关键点的同时预测关键点间的连接概率. 如图3所示,用$\phi_{k_1k_2}$表示两个关键点k_1、k_2的连接概率,两两关键点连接概率集合则表示为

$$\{\phi_{k_1k_2} | (k_1, k_2 \in \eta)\}. \tag{4}$$

其中: η为能被检测到的关键点对; $\phi_{k_1k_2}$定义为

$$\phi_{k_1k_2} = \{p(C|k_1,k_2,t,t+\varepsilon)\}_{\varepsilon \in T}, \tag{5}$$

C为一个二值变量, t为负责检测关键点k_i的网格位置, $t+\varepsilon$为负责检测关键点k_2的网格位置, ε为一个位置偏移量限制在一个连接范围T,是以位置t为原点的$H \times W$范围,即$\varepsilon = (\Delta x, \Delta y)$,且$|\Delta x| < W$, $|\Delta y| < H$. 连接的置信度由真实连接进行监督,训练的优化目标定义为

$$R_{k_1k_2}=$$
$$\min(\delta^i_{k_1},\delta^j_{k_2})\{\delta^i_{k_1}\delta^j_{k_2}-\hat{p}(C|k_1,k_2,t,t+\varepsilon)\}. \quad (6)$$

其中: δ^i_k 用来指示第 i 个网格是否真实负责预测第 k 个关键点; j 表示在连接范围 T 之内负责预测邻近关键点的网格标号; $\delta^i_{k_1}\delta^j_{k_2}$ 表示实际关键点 k_1 与关键点 k_2 能否连接,值为0或1.

基于关键点的检测可得到一些有关各关键点的预测边框信息,这些预测边框对某一个关键点可能存在冗余现象,表现为一个关键点可能被多个预测边框框定,使用标准的非极大值抑制方法过滤掉得分较低的预测边框,剩下得分高的预测边框作为候选的预测边框. 经过非极大值抑制处理后的每个关键点会包含多个候选的预测边框,而要将构成肢体的关键点之间进行匹配和关联,是一个已知的 k 维匹配NP难问题. Cao等[9]采用了能够实时生成一致匹配的松弛方法,该方法首先选择最小的边数构建姿态的生成树骨架,其节点和边分别表示关键点的候选区域子集和它们之间的连接,没有使用完整的图;然后将匹配问题进一步分解为一组偶图匹配子问题来单独确定树的相邻节点匹配等,验证了在极小的计算成本下,采用的贪心推理能够很好地逼近全局解,得出树的非相邻节点之间的关系可以隐式地建模在它们成对的关键点连接概率置信中且可使用卷积网络进行估计的结论. 本文参考了该方法对多人关键点进行匹配和分组.

N_1、N_2 分别表示关键点 k_1 和 k_2 所有候选框的集合,可以定义为所有候选连接集合的最优分配问题,即

$$F_{k_1k_2}=\prod_{N_1}\prod_{N_2}p(C|k_1,k_2,t,t+\varepsilon)^{Z^{n_1n_2}_{k_1k_2}}. \quad (7)$$

其中: $Z^{n_1n_2}_{k_1k_2}$ 为一个二值变量,属于集合

$$Z=\{Z^{n_1n_2}_{k_1k_2}|(k_1,k_2)\in\eta,n_1\in N_1,n_2\in N_2\},$$

表示关键点 k_1 的第 n_1 个候选预测框与关键点 k_2 的第 n_2 个候选预测框能否连接,并且满足以下条件:

$$\forall n_1\in N_1,\forall n_2\in N_2,$$
$$\sum_{N_1}Z^{n_1n_2}_{k_1k_2}=1\bigwedge\sum_{N_2}Z^{n_1n_2}_{k_1k_2}=1. \quad (8)$$

对能够构成连接的每一对 (k_1,k_2) 分别定义偶图匹配子问题,从而找到 k_1 与 k_2 之间连接集的最优分配,得到 k_1 与 k_2 的最优分配

$$Z_{k_1k_2}=\mathrm{argmax}F_{k_1k_2}. \quad (9)$$

最后,通过所有最优分配,可以在多个人中给每人的分组分配各关键点的可连接集合,以组成全身姿态.

2 实验与分析

2.1 实验数据

实验使用两个数据集,其中一个为LSP数据集及其扩充数据集,是单人姿态估计数据集,共包含12 000余张有关体育运动的图像,每张图像只对一人预先进行了标注,每个人在完全可见的情况下有14个关键点标注. 训练时将数据集进行划分,使用9 000张图像作为训练集,剩下的3 000张作为验证集. 另一个为MPII人体姿态数据集,包含24 000余张图像,每张图像已预先进行了标注,图像中包含若干人,每人在完全可见的情况下有16个关键点标注. 训练时将数据集进行划分,使用18 000千张图像作为训练集,剩下6 000余张作为验证集. 在LSP数据集上使用正确关键点百分比(PCK, percentage of correct keypoints)作为衡量标准,在MPII数据集上使用关节点检测的平均精度(mAP, mean average precision)作为评价指标,与多种方法分别进行单人姿态和多人姿态估计对比实验.

2.2 尺度转换结构实验结果

对上采样和所设计的尺度转换结构在3种不同层数的DenseNet上进行对比实验,数据集为MPII多人姿态估计数据集,实验结果如表1所示. 由对比结果可知,尺度转换结构与上采样相比,得到的精度会有些许下降,但速度明显提升,可得知尺度转换结构能够有效降低计算量,从而提高算法运行速度.

表1 上采样与尺度转换结构性能比较

方法	mAP/%	速度/(s/image)
DenseNet-121+上采样	74.7	13
DenseNet-169+上采样	75.6	19
DenseNet-121+上采样	76.1	32
DenseNet-121+STS	74.5	8
DenseNet-169+STS	75.4	14
DenseNet-201+ STS	75.7	21

2.3 单人姿态估计实验结果

将本文基于改进DenseNet网络的方法与几种具有代表性的人体姿态估计方法在单人姿态估计任务上进行准确度和速度的比较,使用LSP数据集对各个方法进行测试. 表2为各方法在设置阈值为0.2时的PCK精度. 图4为各方法平均PCK精度与速度的比较结果.

表 2　各个方法关键点PCK精度比较

方法	头部/%	肩部/%	肘部/%	腕部/%	髋部/%	膝部/%	脚踝/%	平均/%	速度/(s/image)
SHN	97.5	91.7	88.0	83.6	92.3	91.4	85.8	90.0	34
DeeperCut	97.2	94.5	87.3	82.4	86.2	81.7	77.2	86.6	23
LFP	98.1	94.7	92.4	87.2	93.4	92.8	90.1	92.7	51
本文 DenseNet-121	96.3	92.3	87.5	82.8	87.1	82.7	76.2	86.4	5
本文 DenseNet-169	97.9	93.1	88.6	83.2	87.9	84.6	78.1	87.6	9
本文 DenseNet-201	97.4	92.8	89.1	83.5	96.6	84.8	77.4	87.4	12

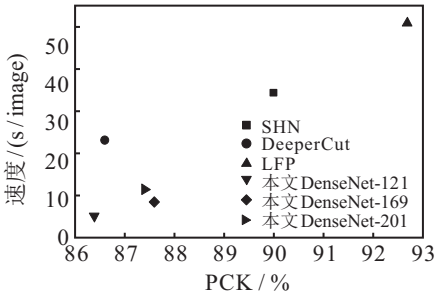

图 4　各方法处理速度与精度关系

方法SHN[5]、DeeperCut[7]和LFP[15]均为自顶向下的方法,其中DeeperCut使用深度为152层的ResNet结构,本文采用深度为121层的DenseNet网络结构,虽然有些关键点的检测精度略低于DeeperCut,但速度优于该方法近3倍. 本文方法与SHN、LFP方法均取得了较高的准确度,这是由于均利用了多个尺度特征信息进行关键点的检测,而SHN和LFP对关键点进行了多个阶段的校准,得到了比单阶段方法更好的精度结果. LEP方法关键点检测平均PCK精度为最高92.7%,该方法通过改进SHN的沙漏结构对多尺度特征信息进行分析推理,使得准确度得到进一步提高,但速度相对慢很多,比改进的DenseNet-121慢了近10倍. 表2中对3种改进的DenseNet网络结构也进行了对比,其中DenseNet-169取得了最好的精度. 在速度与精度之间权衡时,本文方法在保证精度的同时也可将计算成本降低从而加速算法执行,综合两个方面性能,本文方法具有很大优势.

2.4　多人姿态估计实验结果

将本文方法与常用的多人人体姿态估计方法进行精度和速度对比,在MPII数据集进行测试得到各方法对关键点检测的平均精度结果如表3所示. 由表3可见,本文改进的DenseNet-121在头部、肩部和髋部等多个关键点的预测精度要高于其他方法,而脚踝部位的预测精度与ArtTrack[8]方法相差很小,DenseNet-169和DenseNet-201在平均精度上好于其他方法. PAF[9]对关键点进行了多个阶段校准,保证了较好的精度. 本文则是利用了多尺度特征信息,只需进行一次网络推理即可得到更好的精度,其中改进的DenseNet-201比PAF方法mAP高1%. 各方法不同部位的预测精度对比如图5所示. 由图5可以直观看出本文方法整体上优于其他方法.

表 3　各方法的关键点检测平均精度比较

方法	头部/%	肩部/%	肘部/%	腕部/%	髋部/%	膝部/%	脚踝/%	mAP/%	速度/(s/image)
DeeperCut	89.1	83.6	69.6	58.9	68.7	62.6	54.3	69.5	363
ArtTrack	88.7	86.3	75.3	64.2	74.1	68.6	60.8	74.0	6
PAF	90.8	86.9	76.2	66.1	74.2	67.5	61.4	74.7	10
本文 DenseNet-121	91.4	87.1	75.1	64.4	75.6	67.0	60.6	74.5	8
本文 DenseNet-169	92.2	88.7	75.8	65.2	76.3	67.7	61.7	75.4	14
本文 DenseNet-201	92.9	88.2	76.4	65.6	76.4	68.3	62.1	75.7	21

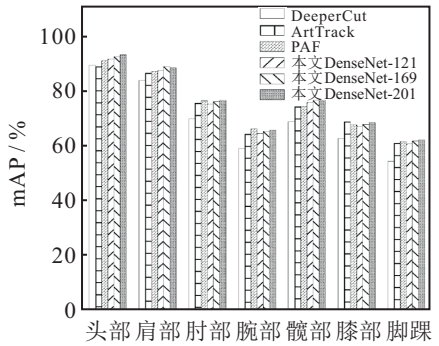

图 5　各部位预测精度比较

同样对各方法进行速度测试. 其中ArtTrack方法处理速度最快,精度也较高,得益于其采用了自顶向下的方式先进行粗略估计,然后使用自底向上的方法进行精细调整,并且简化了身体部位的关系图,从而在速度和精度上均有较好的效果. DeeperCut方法处理速度相对慢很多,并且从表3中可以看出本文方法比DeeperCut方法快近19倍. 随着DenseNet网络层数的增加,精度有所上升,但处理速度会下降,综合速度和精度来看,DenseNet-169的性能最好.

3 结 论

本文针对人体姿态估计任务中人的不同尺度所带来的挑战,提出了一种基于改进DenseNet网络的人体姿态估计方法. 设计一种尺度转换结构,该结构能快速改变特征图尺寸且不会引入附加的参数量和计算量,从而得到多种尺寸的特征图以组成特征金字塔,结合目标检测的思想对关键点位置进行快速检测. 在关键点检测的精度和速度上与其他方法进行了单人姿态估计任务和多人姿态估计任务对比实验,结果表明,相比于其他几种方法,所提出方法具有更高的精度,同时也能保持很快的处理速度,是探索姿态估计处理速度和精度之间均衡性的一种有效方法.

参考文献(References)

[1] Tian Y D, Zitnick C L, Narasimhan S G. Exploring the spatial hierarchy of mixture models for human pose estimation[C]. Proceedings of European Conference on Computer Vision. Berlin: Springer, 2012: 256-269.

[2] Sapp B, Taskar B. MODEC: Multimodal decomposable models for human pose estimation[C]. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 3674-3681.

[3] Toshev A, Szegedy C. DeepPose: Human pose estimation via deep neural networks[C]. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Columbus: IEEE, 2014: 1653-1660.

[4] Wei S E, Ramakrishna V, Kanade T, et al. Convolutional pose machines[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 4724-4732.

[5] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation[C]. Proceedings of European Conference on Computer Vision. Cham: Springer International Publising, 2016: 483-499.

[6] Chou C J, Chien J T, Chen J T. Self adversarial training for human pose estimation[J]. 2017, arViv: 1707.02439.

[7] Insafutdinov E, Pishchulin L, Andres B, et al. DeeperCut: A deeper, stronger, and faster multi-person pose estimation model[C]. Proceedings of European Conference on Computer Vision. Cham: Springer International Publishing, 2016: 34-50.

[8] Insafutdinov E, Andriluka M, Pishchulin L, et al. ArtTrack: Articulated multi-person tracking in the wild[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE, 2017: 1293-1301.

[9] Cao Z, Simon T, Wei S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE, 2017: 1302-1310.

[10] Huang G, Liu Z, Laurens V D M, et al. Densely connected convolutional networks[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE, 2017: 4700-4708.

[11] Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[12] Xia F, Wang P, Chen X, et al. Joint multi-person pose estimation and semantic part segmentation[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii: IEEE, 2017: 6080-6089.

[13] 鞠默然, 罗海波, 王仲博, 等. 改进的YOLO V3算法及其在小目标检测中的应用[J]. 光学学报, 2019, 39(7): 0715004.
(Ju M R, Luo H B, Wang Z B, et al. Improved YOLO V3 algorithm and its application in small target detection[J]. Acta Optica Sinica, 2019, 39(7): 0715004.)

[14] Chu X, Ouyang W L, Li H S, et al. Structured feature learning for pose estimation[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 4715-4723.

[15] 黄继鹏, 史颖欢, 高阳. 面向小目标的多尺度Faster-RCNN检测算法[J]. 计算机研究与发展, 2019, 56(2): 319-327.
(Huang J P, Shi Y H, Gao Y. Multi-scale faster-RCNN algorithm for small object detection[J]. Journal of Computer Research and Development, 2019, 56(2): 319-327.)

[16] Yang W, Li S, Ouyang W L, et al. Learning feature pyramids for human pose estimation[C]. Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice: IEEE, 2017: 1290-1299.

[17] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 779-788.

作者简介

石跃祥(1964−), 男, 教授, 博士, 从事图像处理、智能系统等研究, E-mail: shiyx@xtu.edu.cn;

许湘麒(1994−), 男, 硕士生, 从事图像处理、计算机视觉的研究, E-mail: 635638544@qq.com.

(责任编辑: 郑晓蕾)