文章编号：2095-2163(2023)02-0145-05 中图分类号：TP391.4 文献标志码：A

基于改进沙漏网络的人体姿态估计方法

黄子健，方 宇，杨蕴杰，张伯强，魏旋旋，杨 皓
（上海工程技术大学 机械与汽车工程学院，上海 201620）

摘 要：近年来人体姿态估计已成为计算机视觉领域的热门研究方向，堆叠沙漏网络是人体姿态估计领域中最具代表性的研究成果之一，但该网络对于图像细节特征的提取能力较差。为增强网络对细节特征的处理能力，本文提出了基于改进沙漏网络的人体姿态估计模型。该模型使用 ResNet50 提取高质量的图像底层特征，用步长为 2 的 3×3 卷积核代替 maxpooling 进行下采样，最大程度保留原有图像信息；考虑到不同分辨率下的特征丰富度具有一定差异性，使用不同的残差模块对不同分辨率的 feature map 进行处理，增强网络对特征的学习能力；最后使用反卷积最大化还原原始图像的局部特征。实验结果显示，本文模型在 COCO 测试集上的平均精度达到 74.1%，比堆叠沙漏网络高出 4.7%，检测精度有较大提升。

关键词：堆叠沙漏网络；姿态估计；残差模块；反卷积

Research on human pose estimation based on improved hourglass network

HUANG Zijian, FANG Yu, YANG Yunjie, ZHANG Boqiang, WEI Xuanxuan, YANG Hao
(School of Mechanical and Automotive Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 In recent years, human pose estimation has become a hot research area in the field of computer vision. Stacked hourglass network is one of the most representative research results in the field of human pose estimation, but the network has poor ability to extract image detailed features. In order to enhance the feature processing ability of the network, a human pose estimation model based on improved hourglass network is proposed in this paper. The model uses resnet50 to extract the low-level features of high-quality image and uses 3×3 convolution kernel with stride 2 to replace maxpooling for down sampling to retain the original image information to the greatest extent. Considering the difference of feature richness under different resolutions, different residual modules are used to process feature maps with different resolutions to enhance the learning ability of the network. Finally, deconvolution is used to maximize the local features of the original image. The experimental results show that the average accuracy of the model on the coco test set is 74.1%, which is 4.7% higher than that of the stacked hourglass network.

【Key words】 stacked hourglass network; pose estimation; residual module; deconvolution

0 引 言

人体姿态估计是计算机视觉领域的重要研究方向之一，其目的是在给定的图片或视频中定位人体关节点的位置，进而实现人体姿态的跟踪和预测。姿态估计已被广泛应用于多个领域。如：人机工效学分析[1]、运动分析[2]和康复医疗[3-4]等。人的外观变化（如衣物和体型）和场景的复杂程度，都会对姿态估计结果的准确性产生影响。解决这些问题的关键，是如何利用数据中的特征来提取其中的上下文信息，以及如何利用人体各关节之间的相互位置关系，增强姿态估计结果的准确度。

传统的姿态估计大多使用基于图形结构[5-6]的方法，并依据人体各部件的相对空间位置关系建立人体结构模型；使用方向梯度直方图（Histogram of Oriented Gradient, HOG）[7]和尺度不变特征转换（Scale-Invariant Feature Transform, SIFT）[8]来提取人工设定的特征。虽然该方法在某些场景下能实现较高的检测效率，但在人物遮挡、拍摄角度和图像光照等因素变化时，算法准确性会受到较大影响[9]。

随着人工智能技术的发展，基于深度学习的姿态估计方法因其精度高、泛化能力强等优点，成为人体姿态估计问题研究的主要方法[10]。DeepPose[11]是首个将深度学习用于姿态估计任务的方法，该方法将姿态估计问题转化为图像特征提取和关节点坐标回归问题，使用 AlexNet 提取图像特征，并使用级

基金项目：上海市松江区科技攻关项目（20SJKJGG08C）。
作者简介：黄子健（1996-），男，硕士研究生，主要研究方向：姿态估计；方 宇（1974-），男，博士，教授，硕士生导师，主要研究方向：智能装备与精密控制。
收稿日期：2022-04-22

联回归器修正浅层网络在全连接层回归得到的关节点坐标,提高了算法预测精度。但该方法仅关注关节点坐标信息,忽视了关节点周围的特征信息,导致其预测结果缺乏鲁棒性[12]。Wei 等人[13]提出了卷积姿态机(Convolutional Pose Machines,CPM)模型,CPM 使用一种顺序化的卷积结构来提取图像特征信息,每个阶段将上一阶段的预测结果作为本阶段的输入,输出含有预测关节点信息的热图,并引入中间监督模块,用以解决训练过程中的梯度消失问题。Newell 等人[14]以 CPM 为基础,提出了堆叠沙漏网络(Stacked Hourglass Networks,SHN),该网络由多个沙漏网络堆叠组成。与 CPM 不同的是,SHN 在多个分辨率上学习关节点的特征,并且可以学习关节点之间的结构特征,预测结果更加准确。但是,该网络也存在一些缺陷。由于输入网络的图像仅通过简单预处理,无法提取出更加细致的底层特征用于后续的关节点预测,需要通过增加沙漏模块数量以改善预测精度。此外,因网络的上采样使用最邻近插值方法,在将图像从低分辨率还原为高分辨率的过程中,会丢失大量局部特征信息,影响各关节部位的纹理和形状特征提取。

针对以上问题,本文提出一种改进沙漏网络,主要改进工作为:

(1)使用预训练好的 ResNet50[15]网络取代 SHN 的预处理模块,提高输入网络的图像底层特征的质量,从而提高模型的检测精度。

(2)通过调整残差模块的参数,改进网络中使用的残差模块,使得网络能在不同分辨率下学习到差异化的特征,增强网络对于局部细节特征的学习能力。

(3)使用反卷积替代上采样。相较于最邻近差值方法,反卷积在还原原始图像像素的任务上表现更好,有利于网络对于细节特征的学习。

1 算法描述

1.1 堆叠沙漏网络

沙漏模块的结构(图 1)与全卷积网络(Fully Convolutional Networks,FCN)[16]相似,沙漏模块的前半部分通过 4 个残差模块和下采样操作降低图像分辨率并扩大感受野,在其中心的 C5 卷积层得到最低分辨率和最大感受野的特征图(feature map)后,进行 4 次上采样操作,并将通过跳连接传递的图像特征与上采样得到的图像特征进行融合,逐步将图像还原至高分辨率。SHN 通过堆叠沙漏网络,重复地进行自下而上(高分辨率至低分辨率)、自上而下(从低分辨率到高分辨率)的过程,同时结合中间监督评估整个图像的初始特征和检测结果,学习人体各关节点的图像特征以及各关节点之间的相对位置信息。

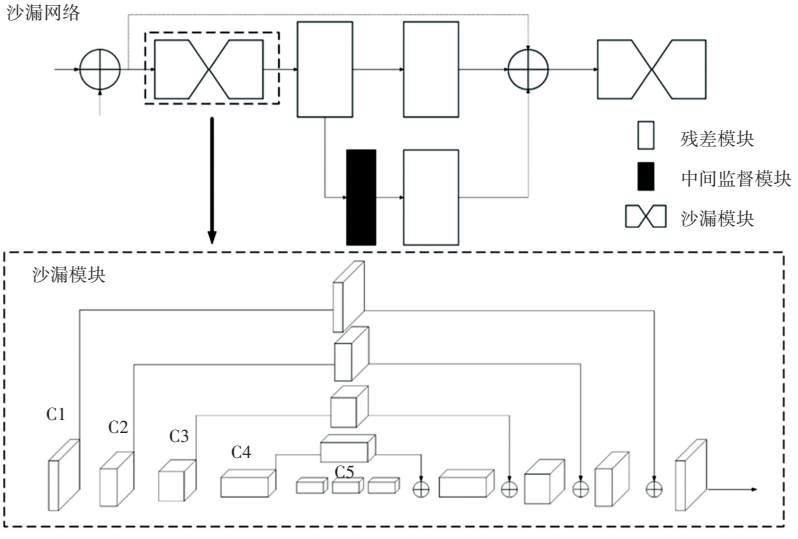

图 1 堆叠沙漏网络

Fig. 1 Architecture of stacked hourglass network

1.2 改进的沙漏网络

1.2.1 预处理模块

SHN 在对图像进行预处理时,首先使用 64 个 7×7、步长为 2 的卷积核,将输入图像的分辨率从 256×256 降至 128×128,随后使用一个残差模块和一次最大池化操作,再将分辨率从 128×128 降至 64×64,以此减少后续训练所需的 GPU 内存量。但是,该操作无法很好地提取图像的底层特征,影响了

SHN 前部沙漏网络的关节点预测精度。因此，使用在 ImageNet 上预训练好的 ResNet50 网络提取更高质量的底层特征，使后续网络能更快地学习到关节点部位的纹理和形状特征。ResNet50 的网络参数见表 1。

表 1　ResNet50 网络参数

Tab. 1　ResNet50 network parameters

layer name	output size	convolution parameters
conv1	$\frac{w}{2}\times\frac{h}{2}$	$7\times7\times64$ 3×3 max pool
conv2_x	$\frac{w}{4}\times\frac{h}{4}$	$\begin{bmatrix}1\times1\times64\\3\times3\times64\\1\times1\times256\end{bmatrix}\times3$
conv3_x	$\frac{w}{8}\times\frac{h}{8}$	$\begin{bmatrix}1\times1\times128\\3\times3\times128\\1\times1\times512\end{bmatrix}\times4$
conv4_x	$\frac{w}{16}\times\frac{h}{16}$	$\begin{bmatrix}1\times1\times256\\3\times3\times256\\1\times1\times1024\end{bmatrix}\times6$
conv5_x	$\frac{w}{32}\times\frac{h}{32}$	$\begin{bmatrix}1\times1\times512\\3\times3\times512\\1\times1\times2048\end{bmatrix}\times3$

由于 ResNet50 输出的图像分辨率为原始图像的 1/32，因此本文使用 3 次反卷积操作，将 ResNet50 输出的图像还原至原图像 1/4 分辨率大小。预处理模块如图 2 所示。

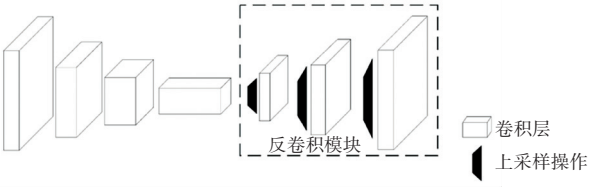

图 2　预处理模块

Fig. 2　Preprocess module

1.2.2　改进的残差模块

SHN 中使用 maxpooling 作为下采样手段，maxpooling 除了可以减少 feature map 的尺寸从而降低计算量外，还能抑制图象中的噪声，被广泛应用于卷积神经网络中。然而，池化会带来特征信息丢失的问题，影响网络预测性能。因此，本文使用步长为 2 的卷积核代替池化实现图片的下采样。在 ResNet 网络中，对于一个如图 3 所示的残差模块，其下采样是在 bottleneck 的 1×1 卷积处完成的，但该方式会使 3/4 的信息丢失，影响下采样得到的 feature map 质量。若推后至在 3×3 卷积处进行下采样操作，使卷积核宽度大于步长，则卷积核在移动过程中能够遍历输入特征上的所有信息，保证下采样后的 feature map 能完整保留原图像特征。

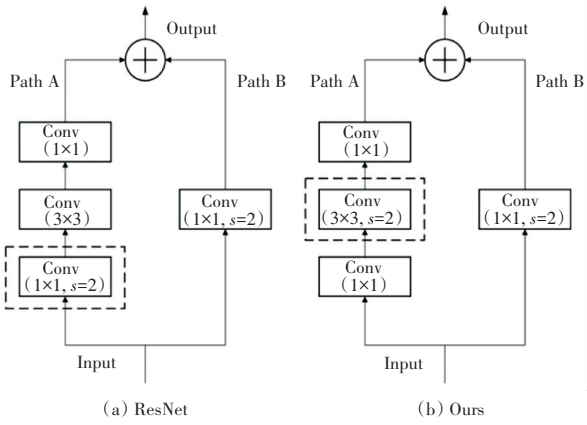

图 3　下采样结构

Fig. 3　Downsample module

当图片输入网络后，随着下采样次数的增加，卷积核的感受野随之增大，在图像下采样至最小辨率时，能得到信息最为丰富的 feature map。由于 SHN 对于不同分辨率的图像均使用相同的残差模块进行处理，并未考虑到不同分辨率下图像信息的差异性。因此，本文使用一种差异化的残差模块，处理不同分辨率下单图像信息。对于不同分辨率的图像，对应的残差模块见表 2。通过使用差异化的残差模块处理不同分辨率的图像信息，能增加网络对图像特征的学习能力，从而提升网络整体性能。

表 2　改进的残差模块参数

Tab. 2　Parameters of improved residual module

layer name	convolution parameters
C1	$\begin{bmatrix}1\times1\times64\\3\times3\times64\\1\times1\times256\end{bmatrix}$
C2	$\begin{bmatrix}1\times1\times64\\3\times3\times64\\1\times1\times256\end{bmatrix}\times2$
C3	$\begin{bmatrix}1\times1\times128\\3\times3\times128\\1\times1\times512\end{bmatrix}$
C4	$\begin{bmatrix}1\times1\times128\\3\times3\times128\\1\times1\times512\end{bmatrix}\times2$
C5	$\begin{bmatrix}1\times1\times256\\3\times3\times256\\1\times1\times1024\end{bmatrix}\times3$

1.2.3 反卷积模块

SHN 使用最近邻差值的方法对图像进行上采样,将图像从低分辨率还原至高分辨率。该方法在还原图像分辨率时,只能根据当前图像的内容进行还原,还原出的图像像素值区分度不强,与下采样前的图像信息差距过大。本文使用反卷积层取代 SHN 中使用的上采样和残差模块,对于每一层反卷积层,记反卷积层输入分辨率为 i,则输出分辨率 o 为

$$o = s \times (i - 1) + k - 2p$$

式中:s 为步幅,k 为卷积核大小,p 为边界扩充参数。

反卷积模块需要按照对应残差模块生成的 feature map 尺寸,适当调整卷积核和通道数的大小,以得到相同分辨率的 feature map。

2 实验结果及分析

为评估本文提出的改进沙漏网络性能,在 COCO[17] 数据集和 MPII[18] 数据集上进行性能验证。本文实验使用 TensorFlow 深度学习框架作为实验框架,操作系统为 Ubuntu18.04 LTS,使用 Python3.6 作为编程语言,GPU 为 RTX3080,显存为 42 GB,采用 Adam 优化器对训练进行优化。

2.1 实验数据及评价标准

2.1.1 实验数据集

本文选用两个公开数据集:COCO 数据集和 MPII 数据集对本文提出的网络进行性能评估。COCO 数据集包括约 200 K 的图像,图像中含有约 250 K 个人体实例。COCO 数据集定义了 17 个人体关键点,并对这些关键点进行了标注。本文将数据集中的 train2017 图像集(约 57 K 张图像,包含 150 K人体实例)作为训练集,val2017 图像集作为验证集。MPII(Max Planck Institute for Information, MPII)人体姿势数据集包含约 25 K 的图像,涵盖 410 种人类日常活动。本文将数据集划分为两部分,其中 20 K 用于训练,其余用于验证。

训练时,使用随机旋转(-30°,30°),随机缩放(0.75,1.25)和左右翻转等数据增强手段,加强网络的泛化能力。

2.1.2 评价标准

对于 COCO 数据集,采用官方指定关节点相似度 OKS(Object Keypoint Similarity, OKS)为模型性能评价的度量方法。OKS 的值在 0~1 之间,越接近 1 说明预测得到的人体关节点与数据集标注的真实值(groudtruth)越相似,预测效果越好。OKS 的定义为

$$OKS = \frac{\sum_i e^{\frac{-d_{pi}^2}{2s_p^2\sigma_i^2}} \cdot \delta(v_{pi} = 1)}{\sum_i [\delta(v_{pi} = 1)]}$$

式中:p 表示 *groudtruth* 中人的 id,i 为每个关键点的 id,d_{pi} 表示每个人预测关键点和 *groundtruth* 的欧氏距离,v_i 为关键点的可见性标志,S_p 为当前人体的尺度因子,σ_i 表示第 i 个关键点的归一化因子,v_{pi} 代表第 p 个人的第 i 个关键点是否可见,δ 用于将可见点选出用于计算。

此外,使用平均精确度 AP(Average Precision, AP)和评价召回率 AR(Average Recall, AR),作为算法在 COCO 数据集上预测性能的评估指标。

对于 MPII 数据集,采用关键点准确估计百分比 PCKh(Head - Normalized Probability of Correct Keypoint, PCKh)作为实验评估指标。若预测的关键点落在 groundtruth 的 αlr(两个参数的乘积)个像素内,则认为该关节点预测正确。其中,α 为一个常数,lr 为参考距离。评估时采用 $\alpha = 0.5$(PCKh@0.5)作为评估标准,lr 取头部边界对角线长度。

2.2 实验结果及分析

本文算法在 COCO 数据集上的结果与其它相关算法所得准确率的对比结果见表 3。可以看出,本算法较 SHN 的 AP 和 AR 分别提升了 4.7%和1.9%。相较于 Baseline[19],本文算法的 AP 和 AR 分别高出 1.2%和 0.5%。

在 MPII 数据集上,测试 7 个部位的准确率。可以看出,本文提出的方法相较其它方法,对于部位检测的准确率更高。

表 3 不同方法在 COCO 数据集上结果对比

Tab. 3 Comparison of different method results in COCO datasets

Method	AP	$AP^{@50}$	$AP^{@75}$	$AP^{@m}$	$AP^{@l}$	AR
DeepPose	66.5	80.6	73.6	63.4	72.7	71.9
CPM	66.9	82.3	76.3	65.3	75.9	74.4
SHN	69.4	85.3	79.8	67.4	78.6	76.8
Baseline	72.9	88.5	80.2	69.0	79.3	78.2
Ours	74.1	88.7	80.1	69.5	78.9	78.7

表 4　不同方法在 MPII 数据集上结果对比

Tab. 4　Comparison of different method results in MPII datasets

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
DeepPose	95.4	94.3	91.7	84.0	89.7	87.0	81.3	89.1
CPM	96.2	95.0	92.0	84.9	90.5	87.7	82.0	89.8
SHN	97.6	95.7	92.3	85.3	91.3	88.6	82.5	90.5
Baseline	97.0	96.5	93.4	88.5	92.0	90.7	83.2	91.8
Ours	97.4	97.7	95.2	89.3	93.1	90.1	85.3	92.6

3　结束语

本文以堆叠沙漏网络为基础进行改进，设计了一种基于改进沙漏网络的人体姿态估计网络。改进模型的预处理模块使用特征提取能力更强的 ResNet50 网络，提高了底层特征的质量；使用改进的残差模块处理不同分辨率的 feature map，并使用步长为 2 的卷积代替最大池化进行下采样操作，确保图像信息的完整性；用反卷积模块扩大图片分辨率，使图片像素值多样化，减少上采样得到的图片与原图片间的信息损失。本文的模型相较 SHN 精度更高，能有效识别遮挡的关节点。如何对模型进行轻量化改进并应用到实际工作中，将是下一步的研究方向。

参考文献

[1] ABOBAKR A, NAHAVANDI D, HOSSNY M, et al. RGB-D ergonomic assessment system of adopted working postures[J]. Applied ergonomics, 2019, 80: 75-88.

[2] 朱洪堃, 殷佳炜, 冯文宇, 等. 一种轻量化实时人体姿势检测模型研究与应用[J]. 系统仿真学报, 2020, 32(11): 2155-2165.

[3] CAPECCI M, CIABATTONI L, FERRACUTI F, et al. Collaborative design of a telerehabilitation system enabling virtual second opinion based on fuzzy logic[J]. IET Computer Vision, 2018, 12(4): 502-512.

[4] 唐心宇, 宋爱国. 人体姿态估计及在康复训练情景交互中的应用[J]. 仪器仪表学报, 2018, 39(11): 195-203.

[5] FISCHLER M A, ELSCHLAGER R A. The representation and matching of pictorial structures[J]. IEEE Transactions on Computers, 1973, 22(1): 67-92.

[6] FELZENSZWALB P F, HUTTENLOCHER D P. Pictorial structures for object recognition[J]. International Journal of Computer Vision, 2005, 61(1): 55-79.

[7] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// International Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2005, 1: 886-893.

[8] LOWE D. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110

[9] 周燕, 刘紫琴, 曾凡智, 等. 深度学习的二维人体姿态估计综述[J]. 计算机科学与探索, 2021, 15(4): 641-657.

[10] CHEN Y, TIAN Y, HE M. Monocular human pose estimation: A survey of deep learning-based methods[J]. Computer Vision and Image Understanding, 2020, 192: 102897.

[11] TOSHEV A, SZEGEDY C. DeepPose: Human pose estimation via deep neural networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1653-1660.

[12] 卢健, 杨腾飞, 赵博, 等. 基于深度学习的人体姿态估计方法综述[J/OL]. 激光与光电子学进展: 1-27[2021-11-28]. http://kns.cnki.net/kcms/detail/31.1690.TN.20210311.1622.003.html.

[13] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 4724-4732.

[14] NEWELL A, YANG K, JIA D. Stacked Hourglass Networks for Human Pose Estimation[C]// European Conference on Computer Vision. 2016: 483-499.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[16] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[17] LIN T, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// European conference on computer vision. 2014: 740-755.

[18] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2d human pose estimation: New benchmark and state of the art analysis[C]// Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. 2014: 3686-3693.

[19] XIAO B, WU H, WEI Y. Simple baselines for human pose estimation and tracking[C]// Proceedings of the European Conference on Computer Vision. 2018: 466-481.