DOI:10.19651/j.cnki.emt.2106017

基于深度学习的自顶向下人体姿态估计算法*

张小娜[1] 吴庆涛[2]
(1.德阳科贸职业学院 广汉 618300;2.西南交通大学 信息科学与技术学院 成都 611756)

摘 要:针对自顶向下的人体姿态估计算法出现的目标框定位错误问题和冗余检测问题,提出了一种基于深度学习的自顶向下人体姿态估计算法。设计了对称空间变换网络与单人姿态估计网络相连接,以从不准确的人体边界框中提出高质量的人体目标框,并且引入了参数化姿态非极大值抑制消除了冗余的姿态估计,应用消除规则对相似的姿态进行消除,得到唯一的人体姿态估计结果。在公共人体姿态估计数据集 MPII 上选取部分数据集进行训练和测试,实验结果表明所提出的方法能够准确地检测出人体关键点,有效地提高了人体姿态估计的准确率,且能够适应人员密集、存在遮挡的复杂场景。

关键词:深度学习;人体姿态估计;对称空间变换网络;姿态非极大值抑制;数据增强
中图分类号:TP391.41 **文献标识码**:A **国家标准学科分类代码**:520.20

Top-down human pose estimation algorithm based on deep learning

Zhang Xiaona[1] Wu Qingtao[2]
(1. Deyang Vocational College of Science and Trade, Guanghan 618300, China; 2. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: Aiming at the problem of object frame positioning error and redundant detection in the top-down human pose estimation algorithm, a top-down human pose estimation algorithm based on deep learning is proposed. The symmetric space transformation network is designed to connect with the single-person pose estimation network to propose high-quality human target frames from inaccurate human body bounding boxes, and parametric pose non-maximum suppression is introduced to eliminate redundant pose estimation, The elimination rule is applied to eliminate similar postures, and the unique human posture estimation result is obtained. Part of the data set is selected for training and testing on the public human pose estimation data set MPII. The experimental results show that the method proposed in this paper can accurately detect the key points of the human body, effectively improve the accuracy of human body pose estimation, and can adapt to crowded people, complex scenes with occlusion.

Keywords: deep learning; human pose estimation; symmetric spatial transformation network; pose non-maximum suppression; data enhancement

0 引 言

随着计算机硬件技术以及深度卷积神经网络的迅速发展[1-3],人体姿态估计技术已成为人体行为分析、步态识别、虚拟现实等领域的热点研究问题。人体姿态估计指的是从单张图像中计算人体各关键点的位置,以形成人体的简易骨架。虽然现在各种人体姿态估计算法相继涌现,但是仍然面临着许多挑战。

从算法的整体框架考虑,目前人体姿态估计算法主要分为两类:自顶向下的人体姿态估计算法和自底向上的人体姿态估计算法,前者首先在图像中检测出人体目标框,然后再在单个人体目标框内进行单人姿态估计[4-6];后者首先在图像中检测出所有人的关键点,再分配到每个人体上[7-9]。自顶向下的人体姿态估计算法由于检测精度较高而受到广大研究学者的青睐,Wei 等[10]设计的顺序卷积结构模型——卷积姿态机构建了多阶段的深度卷积网络,每一阶段都是在前一阶段和原始图像的特征图上进行迭代映射,并再用中级监督进行训练,最终得到了人体各关键点的置信度图。Newell 等[11]提出的堆叠沙漏网络一方面利用多分辨率的热图学习关键点的局部位置特征,另一方面通过多尺度感受野机制学习关键点之间的结构特征,网络的

收稿日期:2021-03-17
* 基金项目:国家自然科学基金(61572406,61976182)项目资助

前后两个部分是对称的，整个网络通过级联的方式堆叠在一起，模型简单且易于扩展。Chen 等[12]提出了级联金字塔结构网络，在单人目标框中通过级联的 GlobalNet 和 RefineNet 实现了对人体关键点的定位，并加入了难关键点挖掘，大幅度提高了人体姿态估计的准确性。

自顶向下人体姿态估计算法虽然准确率较高，但人体关键点定位的准确率容易受人体目标检测算法性能的影响，对于人体目标检测算法出现的检测不准和冗余检测的情况，相对应的单人姿态估计网络分别会出现人体关键点定位错误和人体姿态冗余检测。为了解决上述问题，受文献[13]启发，提出了一种多人姿态估计算法，设计了对称空间变换网络（symmetric spatial transformer network，SSTN）与单人姿态估计网络相连接，以从不准确的人体边界框中提出高质量的人体目标框，并且引入了参数化姿态非极大值抑制消除了冗余的姿态估计，实验结果表明，本文提出的方法有效地提高了人体姿态估计的准确率。

1　自顶向下人体姿态估计算法存在的问题

自顶向下人体姿态估计算法主要存在两个问题：目标框定位错误问题和冗余检测问题。实际上，单人姿态估计网络相当容易受到目标框错误的影响。即使在交并比（intersection over union，IoU）大于 0.5 时将边界框视为正确的情况下，检测到的人体姿态仍然可能是错误的。图 1 所示为因边界框定位错误问题而造成的人体姿态检测错误问题，实线框是标定的真实边界框，虚线框是检测到的 IoU＞0.5 的边界框，对于两种类型的边界框，虚线框的热图中未检测到相应的身体部位。如果将 IoU＞0.5 的虚线框视为检测“正确”的边界框，即使用“正确”的边界框也无法检测到人体姿态。

由于单人姿态估计网络为每个检测出的目标框生成一个姿态，因此冗余检测会导致出现冗余姿态。图 2 所示为冗余检测而造成的冗余姿态问题。

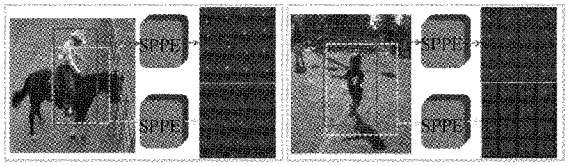

图 1　边界框定位错误问题

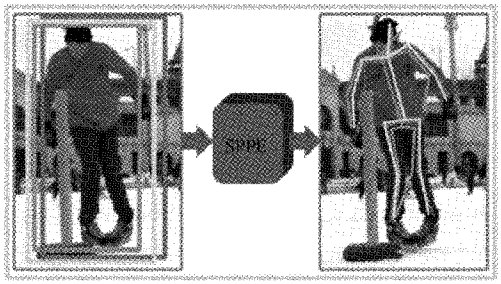

图 2　冗余的人体检测问题

左图显示了检测到的边界框，右图显示了估计的人体姿态。由于自顶向下的人体姿态检测算法都是在每个边界框都是独立进行估计的的，因此可以检测到一个人的多个姿态。

2　多人姿态估计算法

本文提出的多人姿态估计网络结构如图 3 所示，网络主要由对称空间变换网络、并行的单人姿态估计网络和参数化姿态非极大值抑制组成。对称空间变换网络可细分为：空间变换网络（spatial transform network，STN）、单人姿态估计、空间反变换网络（spatial de-transform network，SDTN）。利用人体目标检测算法首先检测出人体目标框的位置，将得到的人体目标框从图像中截取之后放缩到固定的尺度大小，输入到本文的姿态估计网络中生成姿态建议，最后经过参数化姿态非极大值抑制获得最终的人体姿态结果。

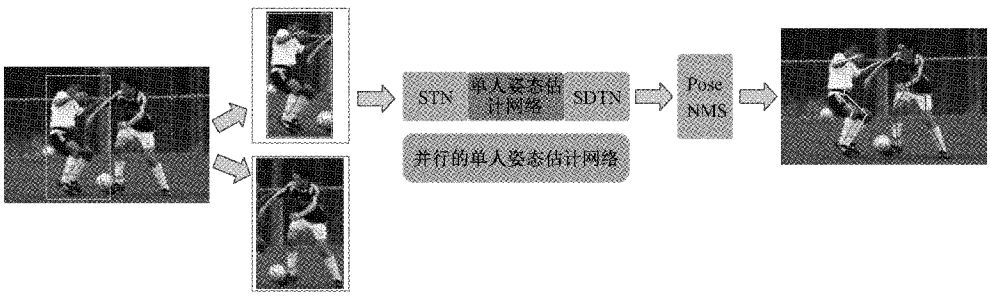

图 3　区域多人姿态估计网络结构

2.1　对称空间变换网络

由于人体目标检测算法提供的人体目标框不够准确，不能很好的适用于单人姿态估计网络，研究表明，人体目标框微小的移动都会严重影响人体姿态估计的结果。本文提出的空间变换网络赋予传统卷积裁剪、平移、缩放及旋转等特性，使得模型具有空间不变性，能够自动的将图像数据进行空间变换以生成一个高质量的人体目标框。

数学上，空间变换网络可以看成二维的仿射变换，可以表示为：

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

其中,θ_1、θ_2、θ_3 是二维向量,$(x_i^s \quad y_i^s)$ 和 $(x_i^t \quad y_i^t)$ 分别为变换前后的坐标。

空间变换网络之后接单人姿态估计网络,用来对人体目标框进行初步的姿态估计,此时,需要将生成的姿态映射到原始的人体目标框图像。自然地,需要空间反变换网络才能将估计的人体姿态重新映射回原始图像坐标。

空间反变换网络通过计算 γ 进行反变换,并基于 γ 生成网格,计算方法可表达为:

$$\begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix} = \begin{bmatrix} \gamma_1 & \gamma_2 & \gamma_3 \end{bmatrix} \begin{pmatrix} x_i^s \\ y_i^s \\ 1 \end{pmatrix} \tag{2}$$

其中,$(x_i^t \quad y_i^t)$ 和 $(x_i^s \quad y_i^s)$ 分别为空间反变换网络变换前后的坐标,由于空间反变换网络是空间变换网络的逆过程,因此可推导:

$$\begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix} = \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}^{-1} \tag{3}$$

$$\gamma_3 = -1 \times \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix} \theta_3 \tag{4}$$

则损失函数 $J(W,b)$ 在空间反变换网络中进行反向传播时,$\dfrac{\partial J(W,b)}{\partial \theta}$ 对于 θ_1、θ_2 可被推导为:

$$\frac{\partial J(W,b)}{\partial \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix}} = \frac{\partial J(W,b)}{\partial \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}} \times \frac{\partial \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}}{\partial \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}} + \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}} \times \frac{\partial \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}}{\partial \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}} \tag{5}$$

对于 θ_3 可被推导为:

$$\frac{\partial J(W,b)}{\partial \theta_3} = \frac{\partial J(W,b)}{\partial \gamma_3} \times \frac{\partial \gamma_3}{\partial \theta_3} \tag{6}$$

$\dfrac{\partial \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix}}{\partial \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}}$ 和 $\dfrac{\partial \gamma_3}{\partial \theta_3}$ 可以分别从式(3)和(4)中推导出来。

在网络训练中,空间反变换网络和单人姿态估计网络一起进行微调。

2.2 并行单人姿态估计网络

为了进一步使得空间变换网络提取到高质量的人体目标框,在训练阶段添加了并行的单人姿态估计网络分支。该分支与原始单人姿态估计网络共享相同的空间变换网络,但是省略了空间反变换网络,此分支的输出直接与位于中心的真实标注的人体姿态标签进行比较并计算损失。在训练阶段,冻结了此并行分支的所有层的权重,其目的是给空间变换网络反向传播中心位置的姿态误差。如果从空间变换网络中得到的姿态不在目标框的中心位置,则并行分支将反向传播误差,这样可以使空间变换网络专注于正确的高质量的人体目标区域。在测试阶段,并行分支将被摒弃。

在训练阶段,可以将并行分支视为调节器。它有助于避免空间变换网络未将姿态转换到提取的人体目标框中心的情况,借助并行单人姿态估计网络,训练的空间变换网络模块可以将人体移动到目标建议区域的中心,以便于通过单人姿态估计网络进行准确的姿态估计。

2.3 姿态非极大值抑制

人体目标检测算法不可避免的会出现冗余检测的情况,这就导致后面的姿态估计网络也随之会出现冗余的姿态估计。为消除多余的姿态估计,本文提出了一种参数化姿态非极大值抑制方法,将置信度最高的姿态估计结果作为参考,应用消除规则对与其相似的姿态进行消除,并重复此过程,直到最后得到唯一剩余的姿态估计结果。

要消除冗余的姿态估计,首先应定义姿态距离度量函数 $d(P_i,P_j \mid \Lambda)$ 来衡量姿态之间的相似度,则消除规则定义为:

$$f(P_i,P_j \mid \Lambda,\eta) = \delta[d(P_i,P_j \mid \Lambda,\lambda) \leqslant \eta] \tag{7}$$

其中,P_i、P_j 分别第 i 个和第 j 个姿态估计结果,Λ 是函数 $d(\cdot)$ 的参数,如果 $d(\cdot)$ 小于阈值 η,则 $f(\cdot)$ 应该输出 1,这表明将姿态 P_j 作为参考姿态的情况下,姿态 P_i 是冗余的,应该被消除。

假设 B_i 为姿态 P_i 的人体目标框,则可定义一个匹配函数为:

$$K_{Sim}(P_i,P_j \mid \sigma_1) = \begin{cases} \sum_n \tanh \dfrac{c_i^n}{\sigma_1} \cdot \tanh \dfrac{c_j^n}{\sigma_1}, & k_j^n \subset \varphi(k_i^n) \\ 0, & \text{其他} \end{cases} \tag{8}$$

其中,k_i^n 表示第 i 个姿态的第 n 个关键点,$\varphi(k_i^n)$ 是在人体关键点 k_i^n 处的边界框,且每个维度是原始目标框 B_i 的 1/10。tanh 函数滤除了具有低置信度的姿态,当两个人体关键点都具有较高的置信度时,输出结果将接近于 1。

此外,还考虑了各人体关键点之间的空间距离,定义为:

$$H_{Sim}(P_i,P_j \mid \sigma_2) = \sum_n \exp\left(-\frac{(k_i^n - k_j^n)^2}{\sigma_2}\right) \tag{9}$$

通过合并式(8)和(9),最终的姿态距离度量函数 $d(P_i,P_j \mid \Lambda)$ 可定义为:

$$d(P_i,P_j \mid \Lambda) = K_{Sim}(P_i,P_j \mid \sigma_1) + \lambda H_{Sim}(P_i,P_j \mid \sigma_2) \tag{10}$$

其中,λ 是衡量两个距离的权重因子,且参数集 $\Lambda = \{\sigma_1,\sigma_2,\lambda\}$。

3 实验及分析

3.1 数据集的建立

本文在公共人体姿态估计数据集 MPII 选取简单场景、人体遮挡、人体密集、复杂姿态等情况下的图像 10 000 张用于网络训练,并选取 1 000 张图像用于对训练的网络模型进行测试。

此外,运用数据增强对选取的数据集进行数据扩充,

以增强网络模型的泛化性。本文在训练网络模型的过程中应用随机水平、垂直翻转，随机旋转范围为±45°，随机缩放图像比例为[0.7，1.35]。通过以上数据增强办法，增加了数据集中样本的复杂度，也可以避免模型的过拟合。

3.2　实验平台和性能评价指标

在 Ubuntu16.04 系统上搭建实验环境，CPU 为 Inter(R) Xeon Silver 4110，GPU 为 NVIDIA GeForce RTX 2080Ti，深度学习框架为 Tensorflow。

选取目标关键点相似度(object keypoint similarity，OKS)来衡量预测的舰载机关键点与真实关键点之间相似度，计算公式为：

$$OKS_p=\frac{\sum_i \exp\{-d_{pi}^2/2S_p^2\sigma_i^2\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)} \tag{11}$$

式中：p 为人体的 ID；i 为人体关键点的 ID，d_{pi} 表示第 p 个人体预测的第 i 个关键点与真实标注的关键点之间的欧氏距离；S_p^2 表示第 p 个人体目标边界框所占的像素面积；σ_i 表示第 i 个关键点归一化因子；v_{pi} 表示第个人体的第 i 个关键点是否可见；δ 为选择函数。

平均准确率(average precision，AP)定义为给定阈值 s，预测的人体关键点结果在整个测试集上的平均准确率，可由测试集所有图片的 OKS 指标计算得到：

$$AP@s=\frac{\sum_p \delta(OKS_p>s)}{\sum_p 1} \tag{12}$$

总体评价指标定义为平均准确率的均值(mean average precision，mAP)，定义为：

$$mAP=mean\{AP@(0.50:0.05:0.95)\} \tag{13}$$

指按给定的阈值 0.5～0.95 按照 0.05 的步长依次计算 AP 后，再取平均值。

3.3　训练和测试

训练过程中误差更新算法选用 Adam，初始学习率设置为 0.000 1，每迭代 20 轮后学习率降为原来的 0.5，小批量数据 batchsize 设置为 8，训练共迭代 80 轮。

测试阶段，选用 SSD[14] 目标检测算法检测人体目标，单人姿态估计网络选用文献[11]中的 SHN 人体姿态估计网络，为确保人体目标检测算法得到的目标框能够覆盖整个人体区域，将检测到的人体目标框高度和宽度方向都扩展了 15%。

为表明本文提出的算法是通用的，再增加一组人体目标检测算法和单人姿态估计网络，分别选用 Faster R-CNN[15] 目标检测算法和文献[12]中的 CPN 姿态估计算法，分别对原算法和加入本文所提算法后的性能进行对比，结果如表 1 所示，带“*”表示加入了本文提出的算法框架。

表 1　不同姿态估计算法对比　%

算法	$AP@0.5$	$AP@0.75$	mAP
SSD+SHN	90.2	79.3	74.1
SSD+SHN*	92.3	82.8	78.6
Faster R-CNN+CPN	91.5	80.6	75.4
Faster R-CNN+CPN*	93.1	82.9	78.8

由表 1 测试的结果可知，加入本文提出的算法后，两组姿态估计算法的平均准确率 mAP 分别提升了 4.5%和 3.4%，表明了本文所提算法的有效性，能够有效提高自顶向下姿态估计算法的准确性，人体姿态估计结果如图 4 所示。

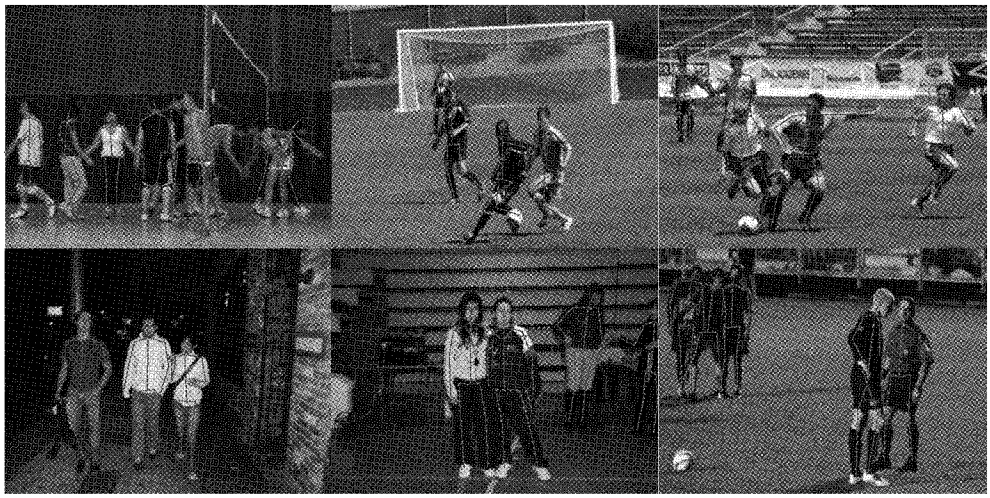

图 4　人体姿态估计结果

由图 4 中的检测结果可知，本文提出的算法由于有效地克服了因人体目标框检测存在偏差和冗余检测而引起的姿态估计错误和冗余姿态的问题，能够准确地检测出人体关键点，且在人员密集、存在遮挡的复杂场景下，算法依

然能够准确的估计出人体姿态。

3.4 消融实验

为了探究所提对称空间变换网络、并行的单人姿态估计网络、姿态非极大值抑制3个方法各自的有效性,以SSD目标检测算法与SHN单人姿态估计算法为例,设计如下实验方案,分别对每组实验方案进行测试,实验结果如表2所示,带"√"表示此实验方案中含有该方法。

表2 消融实验

实验方案	对称空间变换网络	并行单人姿态估计网络	姿态非极大值抑制	*mAP*/%
a				74.1
b	√		√	78.0
c			√	76.4
d	√	√		77.9
e	√	√	√	78.6

对比实验方案a、c、e,可知当缺失对称空间变换网络和并行的单人姿态估计网络时,算法的准确率下降了2.2%,说明提出的对称空间变换网络和并行的单人姿态估计网络的有效性,即可以生成一个高质量的人体目标框用于后续的人体姿态估计;对比实验方案a、b、e,当缺失并行单人姿态估计网络时,算法的准确率下降了0.6%,表明了提出的并行单人姿态的有效性,即可以更好的将人体移动到目标建议区域的中心,以便于通过单人姿态估计网络进行准确的姿态估计;对比实验方案a、d、e,当缺失姿态非极大值抑制时,算法的准确率下降了0.7%,表明提出的姿态非极大值抑制的有效性,即可以有效地消除冗余的姿态。

4 结 论

由于人体关键点定位的准确率容易受人体目标检测算法性能的影响,对于人体目标检测算法出现的检测不准和冗余检测的情况,提出了一种多人姿态估计算法,联合对称空间变换网络与并行的单人姿态估计网络,以从不准确的人体边界框中提出高质量的人体目标框,并且引入了参数化姿态非极大值抑制消除了冗余的姿态估计,实验结果表明,本文提出的方法有效地提高了人体姿态估计的准确率。

参考文献

[1] 丁志敏,邢晓敏,董行,等. 基于深度学习的输电线挂接地线状态目标检测[J]. 电子测量技术,2021,44(3):132-137.

[2] 陶晓天,何博侠,张鹏辉,等. 基于深度学习的航天密封圈表面缺陷检测[J]. 仪器仪表学报,2021,42(1):199-206.

[3] 王永利,曹江涛,姬晓飞. 基于卷积神经网络的PCB缺陷检测与识别算法[J]. 电子测量与仪器学报,2019,33(8):78-84.

[4] 杨兴明,周亚辉,张顺然,等. 跨阶段结构下的人体姿态估计[J]. 中国图象图形学报,2019,24(10):1692-1702.

[5] 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.

[6] PAPANDREOU G, ZHU T, KANAZAWA N, et al. Towards accurate multi-person pose estimation in the wild[C]. CVPR, 2017: 3711-3719.

[7] PISHCHULIN L, INSAFUTDINOV E, TANG S, et al. Deepcut: Joint subset partition and labeling for multi person pose estimation[C]. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1654-1660.

[8] INSAFUTDINOV E, PISHCHULIN L, ANDRES B, et al. Deepcut: A deeper, stronger, and faster multi-person pose estimation model[C]. ECCV, 2017: 3625-3630.

[9] CAO Z, SIMON T, WEI S, et al. Realtime multi-person 2d pose estimation using part affinity fileds [C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2471-2477.

[10] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA: IEEE, 2016: 4724-4732.

[11] NEWELL A, YANG K, AND DENG J. Stacked hourglass networks for human pose estimation[C]. European Conference on Computer Vision, 2016: 483-499.

[12] CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi-person pose estimation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7103-7112.

[13] FANG H S, XIE S Q, TAI Y W, et al. RMPE: Regional multi-person pose estimation[C]. 2017 IEEE International Conference on Computer Vision(ICCV), 2017, 2353-2362.

[14] LIU W, ANGUELOY D, ERHAN D, et al. SSD: Single shot multibox detector [C]. ECCV, Berlin: Springer, 2016: 21-37.

[15] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards realtime object detection with region proposal network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

作者简介

张小娜,硕士,讲师,主要研究方向为软件技术及计算机应用方向。

E-mail: yinjie1020@sina.com

吴庆涛,博士,副教授,主要研究方向为机器学习、集成学习和半监督集成学习等。

E-mail: victorytsb@163.com