JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2021
04
VOL.26

ISSN1006-8961
CN11-3758/TB

细粒度图像分类 *P847*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第4期（总第300期）
2021年4月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑部
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@radi.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial Office of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@radi.ac.cn
Telephone 010-58887035
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

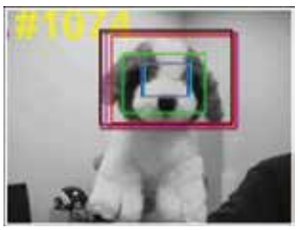
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

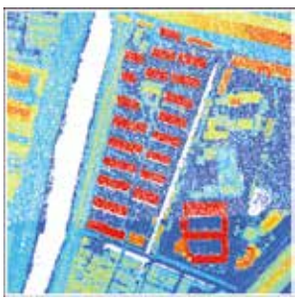
2021年4月 第26卷 第4期（总第300期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

YOLOv3剪枝模型的多人姿态估计(第0837页)



融入时序和速度信息的自适应更新目标跟踪(第0883页)

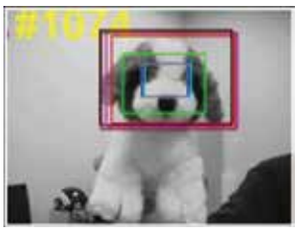


可变半径Alpha Shapes提取机载LiDAR点云建筑物轮廓(第0910页)

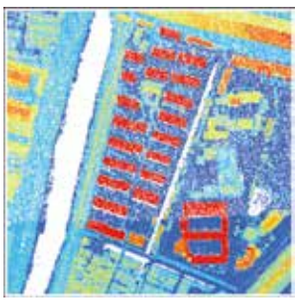
Volume 26, Number 4
Published April 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Research on multiperson pose estimation combined with YOLOv3 pruning model(P0837)



Adaptive update object tracking algorithm incorporating timing and speed information(P0883)



Extraction of building contours from airborne LiDAR point cloud using variable radius Alpha Shapes method(P0910)

中图法分类号:TP391.41　　**文献标识码**: A　　**文章编号**: 1006-8961(2021)04-0837-10

论文引用格式: Cai Z D, Ying N, Guo C S, Guo R and Yang P. 2021. Research on multiperson pose estimation combined with YOLOv3 pruning model. Journal of Image and Graphics,26(04):0837-0846(蔡哲栋,应娜,郭春生,郭锐,杨鹏. 2021. YOLOv3 剪枝模型的多人姿态估计. 中国图象图形学报,26(04):0837-0846)[DOI:10.11834/jig.200138]

YOLOv3 剪枝模型的多人姿态估计

蔡哲栋,应娜*,郭春生,郭锐,杨鹏

杭州电子科技大学通信工程学院, 杭州　310018

摘　要: **目的**　为了解决复杂环境中多人姿态估计存在的定位和识别等问题,提高多人姿态估计的准确率,减少算法存在的大量冗余参数,提高姿态估计的运行速率,提出了基于批量归一化层(batch normalization, BN)通道剪枝的多人姿态估计算法(YOLOv3 prune pose estimator, YLPPE)。**方法**　以目标检测算法 YOLOv3(you only look once v3)和堆叠沙漏网络(stacked hourglass network, SHN)算法为基础,通过重叠度 K-means 算法修改 YOLOv3 网络锚框以更适应行人目标检测,并训练得到 Trimming-YOLOv3 网络;利用批量归一化层的缩放因子对 Trimming-YOLOv3 网络进行循环迭代式通道剪枝,设置剪枝阈值与缩放因子,实现较为有效的模型剪枝效果,训练得到 Trim-Prune-YOLOv3 网络;为了结合单人姿态估计网络,重定义图像尺寸为 256×256 像素(非正方形图像通过补零实现);再级联 4 个 Hourglass 子网络得到堆叠沙漏网络,从而提升整体姿态估计精度。**结果**　利用斯坦福大学的 MPII 数据集(MPII human pose dataset)进行实验验证,本文算法对姿态估计的准确率达到了 83.9%;同时,时间复杂度为 $O(n^2)$,模型参数量与未剪枝原始 YOLOv3 相比下降 42.9%。**结论**　结合 YOLOv3 剪枝算法的多人姿态估计方法可以有效减少复杂环境对人体姿态估计的负面影响,实现复杂环境下的多人姿态估计并提高估计精度,有效减少模型冗余参数,提高算法的整体运行速率,能够实现较为准确的多人姿态估计,并具有较好的鲁棒性和泛化能力。

关键词:目标检测;多人姿态估计;模型剪枝;YOLOv3;堆叠沙漏网络;MPII 数据集

Research on multiperson pose estimation combined with YOLOv3 pruning model

Cai Zhedong, Ying Na*, Guo Chunsheng, Guo Rui, Yang Peng

School of Communication Engineering, Hangzhou Dianzi University, Hangzhou　310018, China

Abstract: **Objective** Estimation of human body posture has always been one of the engaging research directions in computer vision. Attitude estimation in a multiperson complex background is much more difficult than single-person pose estimation (SPPE) in a simple background. Negative factors such as complex background, multiperson recognition, and human occlusion add a large amount of difficulty to the accurate implementation of multiperson pose estimation algorithms. Multiperson pose estimation algorithms can be mainly divided into "top-down" and "bottom-up" frameworks. The essence of the "top-down" framework is from the holistic-local-to-integral process, by detecting the bounding box of the human body and then independently estimating the pose within each frame to complete multiperson pose estimation. The process of the "bottom-up" framework is from the local-to-integral process by first detecting the body parts independently and then assembling the detected body parts into a human body posture. Both frameworks have their own advantages and disadvantages. The use of

收稿日期:2020-05-13;**修回日期**:2020-07-03;**预印本日期**:2020-07-20

*通信作者:应娜　yingna@hdu.edu.cn

基金项目:浙江省自然科学基金项目(LY16F010013)

Supported by: Natural Science Foundation of Zhejiang Province, China(LY16F010013)

a "top-down" framework is susceptible to redundant bounding boxes. The accuracy of pose estimation depends mainly on the quality of the human bounding box. With the "bottom-up" framework, when two or more people are very close together, the gestures that are detected and combined will become very blurred because the framework is localbased and lacks globality. Control is more prone to pose combination errors when applied to multiperson pose estimation in complex environments. We want to complete a more accurate multiperson pose estimation while grasping the overall situation. Therefore, a multi-person pose estimation method combining the YOLOv3 pruning model and SPPE is proposed to solve the problem of positioning and identification of multi-person pose estimation in complex environments, and improve the accuracy of multi-person pose estimation. The YOLOv3 algorithm is a type of end-to-end target detection algorithm proposed in 2018. It uses multiple residual networks for feature extraction and feature pyramid network to achieve feature fusion. Therefore, the YOLOv3 algorithm greatly improves the accuracy of target detection based on maintaining real-time performance. Moreover, the YOLOv3 model has many redundant parameters that greatly affect network operation rate and overall performance. The role of model pruning is to filter the importance of discriminative parameters and remove redundant parameters to reduce the overall model complexity and increase the operating rate. The stacked hourglass network introduced in 2016 consists of multiple hourglass subnets and is extremely malleable. The hourglass subnetwork consists of a residual network that exploits the excellent combining capabilities and feature extraction capabilities of the residual network to extract features of the picture or video. The idea of the primary network-subnetwork provides extremely flexible plasticity for the stacked hourglass network. Multiple subnetwork stacks can help subsequent subnetworks utilize the information extracted by the previous subnetwork, improving the accuracy of the overall network prediction of the human joint points. **Method** The algorithm is based on the target detection algorithm YOLOv3 and the stacked hourglass algorithm. The YOLOv3 network anchor box is modified by the overlap K-means algorithm to adapt to pedestrian target detection better, and the Trimming-YOLOv3 network is trained. In batch normalization, the scaling factor of the layer performs cyclic iterative channel pruning on the Trimming-YOLOv3 network, sets the pruning threshold and scaling factor, achieves a more effective model pruning effect, and trains to obtain the Trim-Prune-YOLOv3 network. To combine the SPPE network, the picture size is redefined to 256 × 256 pixels (nonsquare pictures are implemented by zero padding), then the four hourglass subnetworks are cascaded to obtain a stacked hourglass network, improving the overall attitude estimation accuracy. **Result** This method has been verified by the MPII human pose dataset (MPII dataset) of Stanford University, which is one of the most authoritative datasets in the field of human pose estimation. The MPII is a very challenging multiperson pose dataset, which contains 3 844 training combinations and 1 758 test groups, including occluded people and overlapping people. The MPII dataset contains 16 personal markers, which are the head, shoulders, elbows, wrists, hips, knees, and ankles. On the MPII dataset, the accuracy of the multiperson pose estimation algorithm reaches 83.9%, the time complexity is $O(n^2)$, and the model parameter amount decreases by 42.9% compared with the unpruned original YOLOv3. **Conclusion** The multiperson pose estimation method combined with the YOLOv3 pruning algorithm can effectively reduce the negative effect of complex environments on human pose estimation, achieve multiperson pose estimation, and improve estimation accuracy in complex environments, while using model pruning methods can effectively reduce model redundancy parameters to improve the overall speed of the algorithm. Experimental results show that the method can achieve a more accurate multiperson pose estimation and has better robustness and generalization ability compared with other methods.

Key words: object detection; multi-person pose estimation; model pruning; YOLOv3; stacked hourglass network; MPII dataset

0 引 言

随着计算机软硬件技术的迅速发展,计算机视觉进入了高速发展阶段,人体姿态估计作为热点研究领域之一,同样进入了高速发展期。其中,单人姿态估计算法的较高准确性来源于其高质量的输入信息,但由于多人复杂环境中存在大量不确定性因素和负面影响,如复杂背景、多人识别和人体遮挡等因素给多人姿态估计算法的准确实现增加了很大难度,所以提升复杂环境中的多人姿态估计依旧是一个具有挑战性的工作。

经过多年发展,多人姿态估计算法主要分为“自顶向下”框架和“自底向上”框架。“自顶向下”框架的本质是整体—局部—整体的过程,通过检测人体的包围框,独立估计每个框内的姿态,从而完成多人姿态估计。“自底向上”框架的过程则是局部—整体,先独立检测身体部位,然后将检测到的身体部位组装形成人体姿态。这两种框架各有优缺点,使用“自顶向下”框架容易受冗余的包围框影响,姿态估计的准确性主要取决于人体包围框的质量,这也对包围框的精度有了更高要求。而使用“自底向上”框架,当两个或两个以上的人间距过近时,检测识别到的身体部位通过组合而成的姿态会变得十分模糊,因为这个框架是基于局部性的而缺乏对全局的把控,在应用于复杂环境下的多人姿态估计时更容易出现姿态组合错误。

1 相关工作

在单人姿态估计中,仅通过估计单个人的姿态来简化姿态估计问题。以树模型(Sapp 等,2010;Zhang 等,2009)和随机森林模型(Sun 等,2012a;Dantone 等,2013)为代表的模型已经被证明在单人姿态估计中非常有效。Sun 等人(2012b)广泛研究了一些基于图像的模型,例如 2012 年之后进一步发展的随机场模型(Kiefel 和 Gehler,2014)和图像依赖模型(Hara 和 Chellappa,2013)。随着深度学习在图像检测和目标检测领域的优异表现,将其引入到人体姿态估计也成为了热门的研究方向。Toshev 和 Szegedy(2014)提出了一种基于卷积神经网络(convolutional neural network, CNN)和深度神经网络(deep neural networks, DNN)的 DeepPose 单人姿态估计网络,通过使用级联和局部高精度图像获得更高精度的节点坐标,实现了较高准确性的单人姿态估计。Newell 等人(2016)提出了一种基于卷积神经网络的堆叠沙漏网络(stacked hourglass network, SHN)单人姿态估计网络,可通过多级串联网络达到多尺度识别的效果,提高姿态识别的准确性。

多人姿态估计算法逐渐发展。Gkioxari 等人(2014)提出了一种基于 R-CNN(region-CNN)网络对人体姿态关键点预测和行动分类的算法,在 PASCAL VOC(pattern analysis, statistical modelling and computational learning visual object classes)数据集上对该方法进行了检验,并将其与当前主流的方法进行了对比。Chen 和 Yuille(2015)提出了一种通过图形模型来解析大部分被遮挡的人的方法,该方法将人类模拟为身体部位的灵活组合,以达到多人姿态估计的效果。Pishchulin 等人(2016)提出了 DeepCut 网络,首先检测所有身体部位,然后通过整体线性规划对这些部位进行标记和组装。紧接着 Insafutdinov 等人(2016)提出了一种基于 Deepcut 的 DeeperCut 多人姿态估计算法,通过对节点进行聚类从而判断各个节点属于哪一个人,并对各个节点进行标记,分类属于身体的哪一部分,两者结合输出姿态估计结果,同时,使用残差网络(residual network, ResNet)进行检测提高精度,将丰富的候选节点进行压缩,提高网络的速度和鲁棒性。Iqbal 和 Gall(2016)在目标检测的基础上结合线性规划处理复杂环境下人与人之间的遮挡和拥挤情况,实现较好的多人关键点检测。Fang 等人(2017)利用对称空间变换网络(spatial de-transformer network, SDTN)和目标检测网络获取高质量的人体区域框,从而实现较为优秀的多人姿态估计结果。

国内的多人姿态估计算法中,范佳柔(2018)结合注意力网络和实例分割网络,提出了一种更加鲁棒的自顶向下网络模型,实现多人姿态估计;许忠雄(2018)提出了一种利用双分支深度神经网络同时预测骨点位置和骨点之间空间关系的多人姿态估计算法,并能在不同图像分辨率下均达到实时效果。袁鹏程(2019)利用深度传感器获得人体骨架图,以关节点坐标信息为基础提取特征,提出了一种基于关节点 3 维坐标的人体身份和姿态识别算法;张开军(2019)针对视频目标检测遇到的退化帧问题和人体困难关键点问题,提出了一种基于特征融合的视频多人姿态估计方法,利用相邻帧和增加网络深度来提高视频多人姿态估计的准确性。但同时这些方法也存在局限性,以上多人姿态估计研究方法分别使用了自顶向下框架、自底向上框架和深度传感器获取人体骨架图等方法,研究重点都主要在于提升多人姿态估计的准确性,而对估计算法的运行速率缺乏进一步研究。对于图像识别算法而言,算法的运行速率与准确率同样重要,是考查算法有效性的重要指标。

根据上述研究成果,本文旨在基于全局提高姿

态估计的运行速率,同时完成较为准确的多人姿态估计,提出了基于批量归一化层通道剪枝的多人姿态估计方法(YOLOv3 prune pose estimator, YLPPE)。通过改进后的YOLOv3(you only look once v3)剪枝网络检测识别人体包围框,利用目标切割提取模块切割提取并重定义尺寸,将经过精细化处理后的包围框依次输入堆叠沙漏网络得到人体关节的信息,利用中心点回归方法回归得到多人姿态估计结果。算法主要在3个方面进行了改进:1)在人体目标检测中,首先修改了YOLOv3算法网络模型的锚框个数和锚框尺寸,使生成的锚框更加贴合单类行人目标,并增加多尺度训练用于提升网络模型的鲁棒性;增加目标切割提取模块,将Trimming-YOLOv3网络检测识别的人体包围框重定义尺寸为256×256像素的图像,尽可能避免由于变阔错误可能导致的堆叠沙漏网络估计错误的问题。最终训练得到Trimming-YOLOv3网络。2)在模型剪枝中,设置缩放因子和剪枝阈值均为0.5,利用缩放因子和稀疏性惩罚的批量归一化(batch normalization, BN)层通道剪枝对模型通道进行重要性鉴别,移除冗余通道从而减少模型参数量,这能有效降低模型大小,提高算法运行速率。3)在姿态估计部分,级联4个Hourglass子网络得到堆叠沙漏网络,级联的作用是使后续的Hourglass子网络能够参考之前的Hourglass子网络预测得到的人体关节点信息,从而通过反复提取利用已获得的关节点信息提高人体关节点预测的准确性,得到最终的人体关节点信息预测。实验结果显示,本文算法能够在保持较高检测速率的情况下实现较为优秀的多人姿态估计识别精度。

2 结合YOLOv3剪枝的多人姿态估计研究

本文提出的YLPPE算法流程如图1所示,具体图像实例流程如图2所示。先通过改进的YOLOv3网络得到Trimming-YOLOv3网络,利用模型剪枝方法对Trimming-YOLOv3网络进行通道剪枝,得到Trim-Prune-YOLOv3网络用于检测识别获得人体包围框,为了使堆叠沙漏网络获得较好的输入质量,利用目标切割提取模块将包围框扩大10%切割提取,重定义尺寸为256×256像素(非正方形图像通过补零实现)的图像,并依次输入至堆叠沙漏网络(stacked hourglass networks, SHN)中得到人体关节点信息,之后利用中心点回归方法将人体关节点回归至原始图像并判断人体关节点是否合理,输出最终多人姿态估计建议。

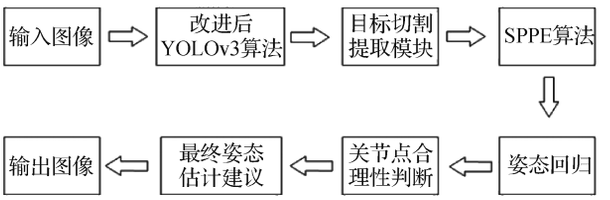


图1 本文算法流程图

Fig. 1 Flowchart of our algorithm

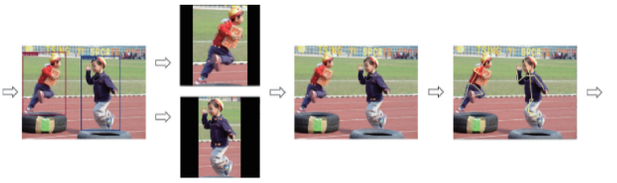


图2 具体图像实例流程

Fig. 2 Specific image flow

2.1 基于重叠度目标维度聚类

YOLOv3算法借鉴了Faster R-CNN(faster regions with CNN features)网络中的锚框(anchor box)思想,用于提高对目标检测分类的能力。但YOLOv3网络的锚框是利用MS COCO(MS common objects in context)数据集的80类目标进行k均值聚类算法聚类得到,并不完全适用于单类的行人目标检测。2维空间的K-means聚类(K-means clustering algorithm)分析一般使用欧氏距离,计算为

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

式中,x_1和x_2是进行聚类的两个点的x轴坐标,y_1和y_2是两个点的y轴坐标,d是指两个点的欧氏距离。欧氏距离适用于点与点之间的距离关系,将距离接近的点分为同一类,但是目标检测领域是用一个包围框而不是一个点来指代目标,所以本文采用重叠度(intersection over union, IOU)进行聚类分析,即候选框与真实框的交集除以并集,消除候选框所带来的误差。重叠度(IOU)计算为

$$I = \frac{T_p}{F_p + T_p + F_n} \tag{2}$$

式中,T_p、F_p和F_n分别表示真阳性、假阳性和假阴性计数。最终的距离函数为

$$d(\boldsymbol{b},\boldsymbol{t}) = 1 - I(\boldsymbol{b},\boldsymbol{t}) \tag{3}$$

聚类目标函数为

$$S = \min \sum_{i=0}^{k} [1 - I(\boldsymbol{b},\boldsymbol{t})] \tag{4}$$

式中，$\boldsymbol{b}$ 为候选框，$\boldsymbol{t}$ 为目标真实框，$I()$ 为重叠度，k 为锚框的个数。通过对算法使用的数据集进行聚类分析，得到适用于本算法的 k 值和锚框长宽值。得到 k 值为 6，即 6 组锚框，各组锚框的长宽具体数值为(2,5)、(10,6)、(16,19)、(30,23)、(32,58)、(62,45)。

其中，算法使用的 ALL-VOC 数据集是利用网上公开的 PASCAL VOC 数据集、MS COCO(Microsoft common objects in context)数据集等综合数据集，筛选提取行人图像和数据得到。由于 VOC 格式的数据集具有易读、易操作和适用广泛等特点，所以将筛选提取的行人图像和数据制作成 VOC 格式的 ALL-VOC 数据集。

2.2　基于 BN 层缩放因子的通道模型剪枝

针对本文目标，在 YOLOv3 的模型中存在着较多冗余参数，这些冗余参数会极大地占用计算量并影响软硬件的运行效率。为了提高参数的有效性并降低模型的整体复杂度，算法利用模型剪枝操作对 YOLOv3 进行剪枝。

本文使用的模型剪枝方法是基于缩放因子和稀疏性惩罚的批量归一化(BN)层通道剪枝。BN 层已经广泛用于 CNN 网络中，作为一种标准方法来使得网络快速收敛并获得更好的性能。BN 层通道剪枝即对批量归一化层输出通道的重要性进行鉴别，移除不重要的通道，从而在保留网络整体准确性的同时大幅降低网络的复杂度。本算法首先给每一个通道都引入 BN 层的缩放因子 γ，然后将缩放因子与通道的输出相乘，接着联合训练网络权重和缩放因子，将缩放因子较小的通道直接移除，并微调剪枝后的网络。剪枝训练的目标函数为

$$L = \sum_{(a,b)} l(f(a,W),b) + \lambda \sum_{\gamma \in \boldsymbol{\Gamma}} g(\gamma) \tag{5}$$

式中，(a,b) 是训练的输入和目标，W 是网络中可训练参数，$\sum_{(a,b)} l(f(a,W),b)$ 是卷积神经网络的训练损失函数，$f(a,w)$ 为针对输入 a 得到的估计值。λ 是平衡因子，$g(\cdot)$ 是缩放因子 γ 的稀疏惩罚项，集合 $\boldsymbol{\Gamma}$ 为 BN 缩放层中已存在的缩放因子。选择的是 L1 正则化用于训练中的自动剪枝，L1 计算为

$$g(\gamma) = |\gamma| \tag{6}$$

本文采用的剪枝方法是一种循环迭代式剪枝，即剪枝—训练—重复的过程，通过判断模型通道的重要性进行多次重复的小规模剪枝，从而避免由于一次性大量剪枝对网络造成不可逆的破坏。具体流程包括：首先对模型权重文件进行稀疏化训练；其次将已经稀疏化的模型进行剪枝操作，若通道的缩放因子小于剪枝阈值则移除该通道；接着将已经剪枝完的网络再次进行稀疏化训练，微调网络；最后，重复上述步骤直至剪枝训练周期结束。完整过程如图 3 所示。

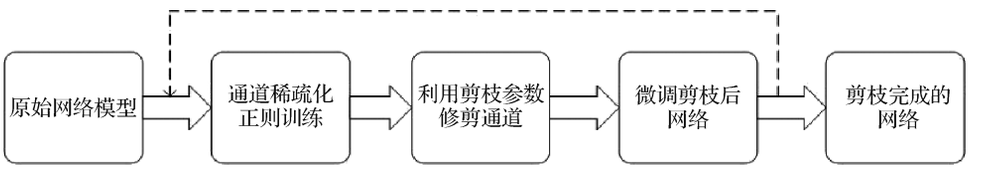


图 3　剪枝流程图

Fig. 3　Pruning flowchart

2.3　多尺度检测训练

利用多尺度检测进行训练能够使网络模型适应不同尺度的输入图像，相比于利用单一尺度训练得到的模型，多尺度检测训练得到的模型具有更优秀的鲁棒性，而且对高分辨率的输入图像检测也有更高的准确率。

YOLOv3 网络中包含大量的卷积层，如果在训练过程中将输入的图像尺寸进行随机改变，利用多尺度输入的方式训练网络模型，可以使网络对不同尺度的图像检测识别具有较好的鲁棒性。在网络的前向传播过程中，张量的尺寸变换是通过改变卷积核的步长来实现的，共经历 5 次缩小从而将特征图缩小到原始输入尺寸的 1/32，具体而言就是在网络的第 1 层、第 5 层、第 12 层、第 37 层和第 62 层卷积层进行特征图的下采样操作从而缩小尺寸。同时在训练过程中随机改变的图像像素分辨率均为 32 的倍数，改变的最小分辨率为 320 × 320 像素，最大为 608 × 608 像素。

2.4 目标切割精细化模块

利用自顶向下框架进行多人姿态估计时,复杂背景和人体遮挡会对姿态估计产生严重影响。为了消除这些可能的负面影响,本文引入目标切割精细化模块。

目标切割提取模块的作用是对利用人体检测算法得到的人体包围框进行合理切割,提取生成新的图像。新的图像是由人体包围框为核心切割得到,原始图像中除了人体目标以外的复杂背景被抛弃,这可以避免若将复杂背景输入至单人姿态估计网络可能导致的错误姿态估计。同时,新生成的图像尺寸符合堆叠沙漏网络输入图像的尺寸要求。为了满足这一功能,同时避免可能由于图像尺寸变形或者目标不完整导致的错误姿态估计,本文通过对图像扩大 10% 切割提取和补零的操作,将包围框切割提取重定义尺寸为 256 × 256 像素(非正方形图像通过补零实现)的图像,图 4 为原始图像识别目标和切割补零后图像。

图 4 原始图像和切割补零后图像

Fig. 4 Original image and post-processing image

2.5 关节点回归及合理性判断

在利用目标切割提取模块降低姿态估计的负面影响后,本文将切割提取后的包围框图像输入至单人姿态估计网络进行人体关节点预测,之后利用中心点回归方法将人体关节点回归至原始图像。

中心点回归方法是指建立以图像中心点为零点的坐标系,得到包围框图像中的人体关节点与零点的位置关系,再将人体关节点回归至原始图像。中心点回归方法具体为:

1)建立以包围框图像中心点为零点的 2 维坐标系,得到人体关节点新坐标 p'_0, p'_1, …, p'_{15},新坐标计算为

$$p'_{xi} = p_{xi} - 128,\ p'_{yi} = p_{yi} - 128$$
$$i = 0,1,2,\cdots,15 \tag{7}$$

式中,包围框图像的尺寸为 256 × 256 像素,p_{xi} 和 p_{yi} 是关节点 p_i 在包围框图像尺寸下的 x 坐标和 y 坐标。

2)将人体关节点新坐标 p'_0, p'_1, …, p'_{15} 回归至原始图像,得到原始图像下的人体关节点坐标 p''_0, p''_1, …, p''_{15},原始图像关节点坐标计算为

$$p''_{xi} = p'_{xi} + c_x,\ p''_{yi} = p'_{yi} + c_y$$
$$i = 0,1,2,\cdots,15 \tag{8}$$

式中,c_x 和 c_y 是原始图像包围框中心点 c 的 x 坐标和 y 坐标。

在利用中心点回归方法完成人体关节点回归后,需要对关节点的合理性进行判断,防止出现由于人体不全造成最终姿态估计混乱的问题。相较于黄铎等人(2019)提出的基于强化学习和 SSD(single shot multi-box detector)算法的多人姿态估计模型,本文提出的 YLPPE 网络模型有较好的鲁棒性,即使出现人体部分缺失的情况仍能有较好的识别结果,并且会将缺失部分的关节点回归至包围框左上角的左上侧。如图 5 所示,图 5 右侧目标双脚的脚踝不在图像中,所以本文算法会将这两个关节点回归至包围框左上角的左上侧,这样在判断人体回归关节点的有效性时可以较为明显地降低错误率,减小姿态估计误差。

图 5 人体不全问题

Fig. 5 Missing body problem

关节点合理性判断的作用是改善由于人体不全造成的姿态估计混乱问题。在通过中心点回归方法将人体 16 个关节点回归至原图后,需要将 16 个关节点坐标与人体包围框左上角坐标进行对比,忽略大于包围框左上角坐标的人体关节点。

3　实验过程

3.1　数据集及实验环境

本文提出的 YLPPE 算法使用了两个数据集：

1) 在目标检测部分，使用的是筛选 PASCAL VOC 数据集和 MS COCO 数据集中的行人图像和数据制作得到的 ALL-VOC 数据集。网上公开的数据集基本都属于综合数据集，行人只是其中很小的一部分，而其他类别的数据对本文算法的训练与使用没有任何帮助甚至可能产生负影响，所以制作只包含行人的数据集尤为重要。利用网上公开的 VOC 数据集、MS COCO 数据集等综合数据集，筛选提取行人图像和数据。由于 VOC 格式的数据集具有易读、易操作和适用广泛等特点，所以将筛选提取的行人图像和数据制作成 VOC 格式的 All-VOC 数据集。在这个数据集中，训练验证集约占整个数据集的 50%，训练集和验证集各占训练验证集的 50%，测试集约占整个数据集的 50%，共 237 088 幅图像。

2) 在人体姿态估计部分，使用的是 MPII 数据集。MPII 多人数据集是一个非常具有挑战性的多人姿态数据集，由 3 844 个训练组和 1 758 个测试组组成，其中包括了被遮挡和重叠的人。

3.2　训练参数的设置

在目标检测部分，Trimming-YOLOv3 网络训练使用的是 DarkNet 深度学习框架，该框架基于随机梯度下降(stochastic gradient descent, SGD)算法来实现网络模型的训练。在满足硬件条件的基础上，为了使模型取得最好的效果，设置训练参数 batch_size 为 64，批次拆分值 subdivisions 为 8，这两个参数的意义是训练时若无法将 64 幅图像同时输入网络，那么就分 8 次、每次 8 幅图像输入网络进行训练。在保存模型方面，设置为每 5 000 次保存一次模型。使用变学习率策略训练 Trimming-YOLOv3 网络模型时，为了防止训练时出现由于学习率过大产生梯度爆炸问题，设置初始学习率 learning_rate 为 0.000 1，学习率衰减参数 scales 为 0.1，变学习率系数 steps 为 45 000 和 90 000。设置最大迭代次数 max_batches 为 125 000。并利用多尺度训练的方法进行网络模型训练，从而起到数据增强的作用，使网络模型具有更好的鲁棒性。

因为原始 YOLOv3 算法训练中的锚框参数是对应 MS COCO 数据集中的 80 类目标设置的，本文算法只需要针对行人这一个目标即可，所以如 2.1 节所述，根据式(2)—(4)训练得到基于实验所用数据集的锚框为(2,5)、(10,6)、(16,19)、(30,23)、(32,58)、(62,45)。

YOLOv3 的剪枝训练是一个循环迭代的过程，在每一次迭代的过程中裁剪部分通道，循环迭代以尽可能达到较好的模型结果。剪枝训练时基于 Pytorch0.4 深度学习框架，经过实验，设置完整训练周期 epoch 为 100，缩放稀疏率 s 为 0.000 1，剪枝阈值为 0.5，BN 层缩放因子 η 初始为 0.5。

在人体姿态估计部分，堆叠沙漏网络基于动量法(momentum)来实现网络模型的训练。设置训练的 epoch 次数为 90、学习率为 0.000 25、变学习率参数 schedule 为 69 和 90、变学习率参数 gamma 为 0.1，同时设置 flip 参数使数据在输入时进行选择以提高模型的鲁棒性。

3.3　实验结果与分析

Trimming-YOLOv3 网络训练过程中的损失(loss)变化情况如图 6 所示，图中横轴为训练批次 batches，纵轴为训练损失值 loss。可以看到损失逐渐收敛，迭代达到 20 000 次时降低至 0.4，然后损失曲线逐渐平缓，在 50 000 次时降低至 0.3，在 60 000 次左右趋于稳定，最终损失值稳定在 0.2 左右。

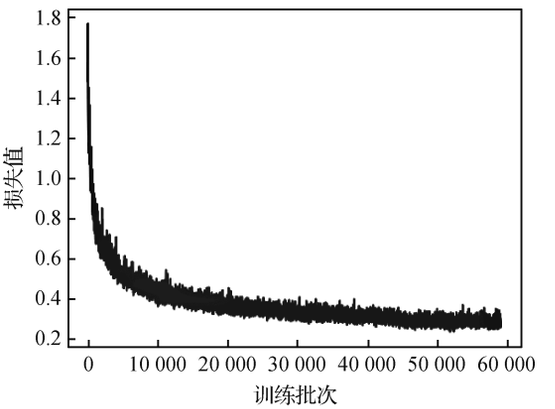


图 6　Trimming-YOLOv3 算法训练损失变化曲线

Fig. 6　Training loss curve of Trimming-YOLOv3 algorithm

Trim-Prune-YOLOv3 网络和 Trimming-YOLOv3 网络及原始 YOLOv3 网络在重叠度(IOU)、召回率(recall)和精度(accuracy)上的比较见表 1。

Trim-Prune-YOLOv3 网络除了在召回率上略微低于原始YOLOv3网络外，在重叠度和精度上都明

表 1　Trim-Prune-YOLOv3 算法与 Trimming-YOLOv3 算法及原始 YOLOv3 算法的比较

Table 1　Comparison of Trim-Prune-YOLOv3 algorithm with Trimming-YOLOv3 algorithm and original YOLOv3 algorithm

	原始 YOLOv3 算法	Trimming-YOLOv3 算法	Trim-Prune-YOLOv3 算法
重叠度	0.756	**0.795**	0.789
召回率	**0.971**	0.968	0.962
精度	0.945	**0.956**	0.953

注：加粗字体为每行最优值。

显优于原始算法；Trim-Prune-YOLOv3 网络在 3 个参数上都略低于 Trimming-YOLOv3 网络，但是相较于模型体积的减少和算法复杂度的降低，这些略微的参数下降是在可接受范围之内的。

在 NVIDIA Titan X 显卡加速下，剪枝后 YOLOv3 网络的整体参数量、模型体积、运行速率和精度与相同条件下剪枝前 YOLOv3 网络的比较如表 2 所示。可以发现整体网络参数下降 42.9%，模型体积下降至 135.11 MB，前向推断耗时下降也较为明显，但精度并没有较大损失。

表 2　Trim-Prune-YOLOv3 算法与 Trimming-YOLOv3 算法网络参数比较

Table 2　Comparison of network parameters between Trim-Prune-YOLOv3 algorithm and Trimming-YOLOv3 algorithm

	Trimming-YOLOv3 算法	Trim-Prune-YOLOv3 算法
参数量/MB	59.6	**34.0**
模型体积/MB	236.6	**135.11**
前向推断/ms	18.2	**11.2**
精度	**0.956**	0.953

注：加粗字体为每行最优值。

堆叠沙漏网络训练过程中的精度值和损失值变化情况如图 7 和图 8 所示。图 7 可以看到随着训练的进行，精度值最终稳定在 88% 左右，而损失在训练初期就急剧收敛接近于 0.01，最终稳定在 0.01 左右。

本文算法在 MPII 数据集上进行了验证，完整的精度结果参见表 3，其中，RMPE 为 regional multi-person pose estimation（Fang 等，2017）。通过与多个多人姿态估计算法的精度比较可以发现，本文算法在人体关节点检测上表现优秀，在整体的精度上处于最优。

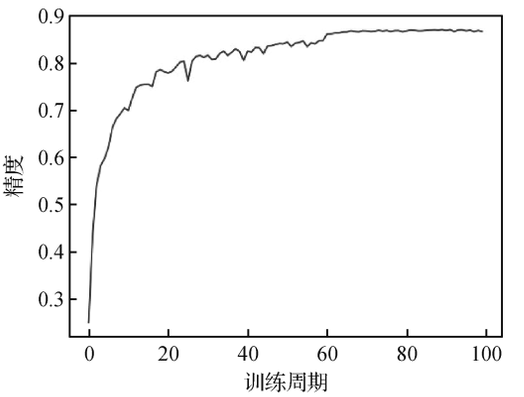


图 7　堆叠沙漏网络训练精度变化曲线

Fig. 7　Training accuracy curve of stacked hourglass network

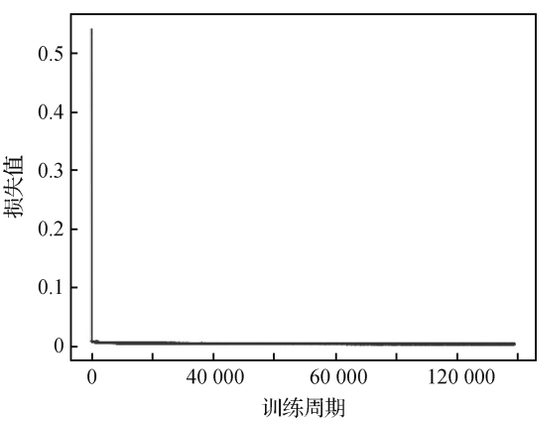


图 8　堆叠沙漏网络训练损失变化曲线

Fig. 8　Training loss curve of stacked hourglass network

表 3　MPII 数据集验证结果（mAP）

Table 3　Validation results in the MPII dataset（mAP）

/%

	Iqbal 和 Gall（2016）	DeeperCut（Insafutdinov 等，2016）	RMPE（Fang 等，2017）	本文
头	58.4	78.4	**88.4**	76.2
肩	53.9	72.5	86.5	**88.1**
肘	44.5	60.2	78.6	**87.1**
腕	35.0	51.0	70.4	**87.4**
臀	42.2	57.2	74.4	**82.8**
膝	36.7	52.0	73.0	**81.6**
踝	37.1	45.4	65.8	**83.2**
整体	43.1	59.5	76.7	**83.9**

注：加粗字体为每行最优值。

测试集中随机选择多幅图像进行测试，结果如图 9 所示。从图 9 中可以发现，本文算法对姿态估计有较高的准确率，在小目标的检测上表现优异。但是也可以看到，当图像中的人体被过多遮挡或者间杂人体躯干遮挡时，算法准确性会明显下降，主要是由于自顶向下框架本身存在的部分问题以及数据集中这类图像的数据标签缺失导致的。最后，在实时性上，本文 YLPPE 算法在 Titan X 显卡加速的条件下对视频流的检测速度达到 39 帧/s 左右，相较于基于剪枝前算法的多人姿态估计算法 24 帧/s 的速度有较为明显的提升，但仅降低了 0.2% 的错误率，表明了剪枝的有效性。

图 9　YLPPE 算法效果演示图

Fig. 9　YLPPE algorithm effect demo

4　结　论

在现有多人姿态估计方法的基础上提出了一种结合 YOLOv3 剪枝的多人姿态估计算法。本文算法通过对 YOLOv3 网络进行合理改进和模型剪枝，再通过结合堆叠沙漏网络成功实现了多人姿态估计的效果，在精度和速度上相较于其他算法有较为明显的提升，并且在 MPII 数据集上验证了可行性。实验结果表明，本文提出的 YLPPE 算法在 MPII 精度上可以达到 83.9% 的精度，在 Titan X 显卡加速下，视频流的检测速度在 39 帧/s 左右。与主流的多人姿态估计相比，本文算法在识别精度上处于较优位置，同时利用剪枝方法降低了模型体积和算法复杂度，进一步提高了检测速率。但本文算法依旧存在较多不足，如受限于数据集的图像类型，对半身及更少人体躯干的检测性能较差；图像视频中人体遮挡情况较为严重的场景下，识别精度不理想等。

多人姿态估计研究的重点及难点在于平衡和兼顾识别精度和速度，本文利用剪枝方法实现了提升识别速度的目的，但如何提升识别精度依旧是后续研究的重要方向。在未来工作中，优化网络结构是有效解决上述问题的一个研究方向，这能够从根本上提升多人姿态估计的识别精度和速度；针对行人重叠遮挡较为严重时的姿态估计问题，除了在网络结构上进行改进，也可以从数据集的补充扩展方向开展后续研究工作，从而帮助算法具有更好的整体性能。

参考文献（References）

Chen X J and Yuille A. 2015. Parsing occluded people by flexible compositions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3945-3954 [DOI: 10.1109/CVPR.2015.7299020]

Dantone M, Gall J, Leistner C and Van Gool L. 2013. Human pose estimation using body parts dependent joint regressors//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 3041-3048 [DOI: 10.1109/CVPR.2013.391]

Fan J R. 2018. Multi-Person Pose Estimation Based on Deep Learning. Hangzhou: Zhejiang University (范佳柔. 2018. 基于深度学习的多人姿态估计. 杭州：浙江大学)

Fang H S, Xie S Q, Tai Y W and Lu C W. 2017. RMPE: regional multi-person pose estimation//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2353-2362 [DOI: 10.1109/ICCV.2017.256]

Gkioxari G, Hariharan B, Girshick R and Malik J. 2014. Using k-poselets for detecting people and localizing their keypoints//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 3582-3589 [DOI: 10.1109/CVPR.2014.458]

Hara K and Chellappa R. 2013. Computationally efficient regression on a dependency graph for human pose estimation//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 3390-3397 [DOI: 10.1109/CVPR.2013.435]

Huang D, Ying N and Cai Z D. 2019. Optimization of estimation algorithm for the multi-person pose based on reinforcement learning. Computer Applications and Software, 36(4): 186-191 (黄铎, 应娜, 蔡哲栋. 2019. 基于强化学习的多人姿态检测算法优化. 计算机应用与软件, 36(4): 186-191) [DOI: 10.3969/j.issn.1000-386x.2019.04.029]

Insafutdinov E, Pishchulin L, Andres B, Andriluka M and Schiele B. 2016. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 34-50 [DOI: 10.1007/978-3-319-46466-4_3]

Iqbal U and Gall J. 2016. Multi-person pose estimation with local joint-to-person associations//Proceedings of 2016 European Conference on Computer Vision. Amsterdam, The Netherlands: Springer: 627-642 [DOI: 10.1007/978-3-319-48881-3_44]

Kiefel M and Gehler P V. 2014. Human pose Estimation with fields of parts//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 331-346 [DOI: 10.1007/978-3-319-10602-1_22]

Newell A, Yang K Y and Deng J. 2016. Stacked hourglass networks for human pose estimation//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 483-499 [DOI: 10.1007/978-3-319-46484-8_29]

Pishchulin L, Insafutdinov E, Tang S Y, Andres B, Andriluka M, Gehler P and Schiele B. 2016. DeepCut: joint subset partition and labeling for multi person pose estimation//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4929-4937 [DOI: 10.1109/CVPR.2016.533]

Sapp B, Toshev A and Taskar B. 2010. Cascaded models for articulated pose estimation//Proceedings of the 11th European Conference on Computer Vision. Crete, Greece: Springer: 406-420 [DOI: 10.1007/978-3-642-15552-9_30]

Sun M, Kohli P and Shotton J. 2012a. Conditional regression forests for human pose estimation//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 3394-3401 [DOI: 10.1109/CVPR.2012.6248079]

Sun M, Telaprolu M, Lee H and Savarese S. 2012b. An efficient branch-and-bound algorithm for optimal human pose estimation//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 1616-1623 [DOI: 10.1109/CVPR.2012.6247854]

Toshev A and Szegedy C. 2014. DeepPose: human pose estimation via deep neural networks//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1653-1660 [DOI: 10.1109/CVPR.2014.214]

Xu Z X. 2018. Research and Application of Real Time Multi-person Pose Estimation for Surveillance Video. Wuhan: Wuhan University (许忠雄. 2018. 监控视频实时多人姿态估计算法研究与应用. 武汉: 武汉大学)

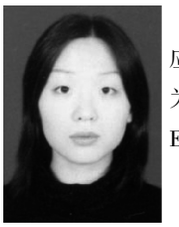
Yuan P C. 2019. Research on Human Body Gesture Recognition Algorithm Based on Convolutional Neural Network. Xi'an: Xi'an Shiyou University (袁鹏程. 2019. 基于卷积神经网络的人体姿态识别算法研究. 西安: 西安石油大学)

Zhang K J. 2019. Key Technology and Application of Visual Object Detection and Recognition Based on Deep Learning. Nanjing: Nanjing University (张开军. 2019. 基于深度学习的视觉目标检测与识别关键技术及应用. 南京: 南京大学)

Zhang X Q, Li C C, Tong X F, Hu W M, Maybank S and Zhang Y M. 2009. Efficient human pose estimation via parsing a tree structure based human model//Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE: 1349-1356 [DOI: 10.1109/ICCV.2009.5459306]

作者简介

蔡哲栋,1994 年生,男,硕士研究生,主要研究方向为图像处理。
E-mail: 472915093@qq.com



应娜,通信作者,女,副教授,主要研究方向为信号处理、语音信号处理及其应用。
E-mail: yingna@hdu.edu.cn

郭春生,男,副教授,主要研究方向为视频分析与模式识别。
E-mail: guo.chsh@gmail.com
郭锐,男,副教授,主要研究方向为无线通信、信道编码、信源信道联合编码。E-mail: guorui@hdu.edu.cn
杨鹏,男,硕士研究生,主要研究方向为图像处理。
E-mail: 394804791@qq.com