2020 年 10 月 第 10 期
Vol. 41 No. 10 2020

轻量级双路卷积神经网络与帧间信息推理的人体姿态估计

陈昱昆，汪正祥，于莲芝

（上海理工大学 光电信息与计算机工程学院，上海 200093）
E-mail: chenyukunin1906@163.com

摘 要： 为了提高视频中人体姿态估计检测效果，在保留结构化信息的同时弥补下采样导致的空间分辨率的损失，增加视频中检测效率，本文结合时序信息提出了一种轻量级双路神经网络帧间信息推理的视频人体姿态估计方法. 首先，基于最新的人体关键点检测网络训练一个基于该方法两路融合全卷积网络，一路选用金字塔全卷积网络，并选用采用轻量级 Inverted residuals 作为网络模块，另外一路保持分辨率大小不变以减少空间分辨率的损失，然后提出了一种利用帧间关键点信息建立时序模型，从而推理预测帧的关键点信息. 本文在 PoseTrack 数据集中与最新的方法进行比较，关键点检测 mAP 提高 1.3%，速度提升 20%，关键点跟踪 MOT 提高 2.8%，经过实验验证，本文算法可以保留结构化信息的同时有效弥补空间分辨率的损失并提高检测精度，同时提高了视频检测中的速度.

关 键 词： 姿态估计；关键点检测；关键点跟踪；双路全卷积网络；轻量级；帧间信息
中图分类号： TP389 **文献标识码：** A **文 章 编 号：** 1000-1220(2020)10-2219-06

Human Pose Estimation Based on Lightweight Two-way Convolutional Neural Network and Inter-frame Information Reasoning

CHEN Yu-kun, WANG Zheng-xiang, YU Lian-zhi
(School of Optical-Electrical and Computer Engineering, University, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In order to improve the detection effect of human pose estimation in video, make up for the loss of spatial resolution caused by down sampling while retaining structured information, and increase the detection efficiency in video, this paper proposes a method of human pose estimation in video based on the frame information inference of lightweight two-way neural network combined with temporal information. First of all, based on the latest human key detection network training, a two-way fusion full convolution network based on this method is trained, one is pyramid full convolution network, and the other is light-weight inverted residuals as the network module, and the other keeps the resolution size unchanged to reduce the loss of spatial resolution. Then, a time sequence module is proposed by using the information of key points between frames And then infer the key information of the prediction frame. In this paper, compared with the latest methods in Posetrack data set, the key point detection map is increased by 1.3%, the speed is increased by 20%, and the key point tracking MOT is increased by 2.8%. Through experimental verification, the algorithm in this paper can retain structured information while effectively making up for the loss of spatial resolution and improving the detection accuracy, while improving the speed in video detection.

Key words: attitude estimation; key detection; key tracking; two-way full convolution network; lightweight; inter frame information

1 引 言

近年来，由于注意力机制的发展和深度视觉表征的迅速崛起，视觉理解，如物体检测和场景识别[1-4]等任务已经有了显著的发展. 人体姿态估计[5,6]是许多计算机视觉应用中的重要组成部分，常常应用在视频监控和体育视频分析等领域. 随着深度卷积神经网络的发展，人体姿态估计已经取得了很大突破，但是由于人体遮挡，图片背景复杂以及人体在图片中尺度的不同等因素，仍然是一个具有挑战性的项目. 在视频图像人体姿态估计中，运动物体可能会导致图像模糊，对姿态估计造成较大的困难，同时由于常用神经网络结构比较复杂以及参数较多，导致模型速度一般较慢，因此在视频图像中，对人体姿态估计提出了更高的要求.

单人姿态估计已被广泛研究. 传统方法包括使用梯度方向直方图和可变形零件模型等，这类方法比较依赖于手工定义特征，表达能力有限，很难扩展到复杂的场景. 关于人体姿态估计的研究近年开始从经典方法转向卷积神经网络. Toshev 等人[7]，使用卷积神经网络直接回归人体关键点的坐标，Newell 等人[8]，将回归坐标变成由坐标生成的关键点热力图为标签，并依据深度神经网络采用“下采样上采样”的架构，中间使用跳过层连接，该架构简称为沙漏网络. 这种结构通过下采样对不同特征进行融合和提取，通过上采样与低维特征进行融合，得到原图大小的预测热力图. 这种结构高分辨率表征主要是由低分辨率表征通过上采样得到的高分辨率表

收稿日期：2019-11-18 收修改稿日期：2020-01-07 基金项目：国家自然科学基金项目(61603257)资助. 作者简介：陈昱昆，男，1994 年生，硕士研究生，研究方向为计算机视觉、人体姿态估计；汪正祥，男，1962 年生，硕士，讲师，研究方向为机器视觉；于莲芝，女，1966 年生，博士，副教授，研究方向为计算机视觉、模式识别等.

征,其本身虽然拥有很好的语义表达能力,但是上采样本身并不能完整地弥补空间分辨率的损失.

与单人姿态估计相比,多人姿态估计需要解析场景中所有人的全身姿势,由于多人之间的遮挡,不同关键点和人之间的相互作用使其成为更具有挑战性的任务. 多人姿势估计方法可以分为两类: 分别简称"自上而下"和"自下而上". "自上而下"的方法[6-10]主要将人体关键点分为两个阶段,首先使用性能良好的人物检测器对图像中的人物进行检测,然后对检测出的每个人进行姿态估计,这样做的目的将多人姿态估计转化为单人姿态估计,这类方法非常依赖于人体探测器的性能. CPN[10]介绍了 COCO[11] 2017 年的人体关键点挑战赛胜利者的方法,采用改进的金字塔网络 FPN[12] 作为人体检测器,同时人体关键点检测分为两个阶段,GlobalNet[10] 是使用金字塔网络识别简单的点,RefineNet[10] 采用难点挖掘方法(OHEM)识别困难的关键点并只返还困难点的梯度,进而训练识别困难的关键点. 但这种方法,更容易检测形状小的行人,同时 recall 比较高,但是由于经过行人检测化多人关键点检测为单人关键点检测,两个深度学习模型速率一般较慢. "自下而上"的方法首先预测图像中所有身体关节,然后利用矢量图或者聚类的方法将这些点进行分类,构成不同人的完整姿态. Openpose[13] 是基于"自下而上"方法由卡内基梅隆大学开源的人体关键点检测项目,首先找到图像中的所有人的关键点,然后再对这些点进行匹配连接,使得同一个人的点进行相连,获得最终的姿态估计结果. 这种方法由于只经过一个深度学习模型,时效性比较强,但没有使用行人检测器,而是依赖于语义信息和关键点之间的关系,对关节关系进行建模可能并不那么可靠,准确率比较低,有时无法区分该点属于哪个人,所以本文采用"自上而下"方法,并选用轻量级网络作为骨架网络,在保证准确率的同时提高模型时效性.

上述多人姿态估计,主要是针对单幅图像处理,而在应用场景往往是以视频的形式呈现. 运动物体导致图像模糊以及姿态估计器比较耗时,这对视频检测提出了更高的要求. 视频具有连贯性,视频的帧间信息具有相似性和时序性,如果可以不通过姿态估计器而是通过帧间信息推理出预测帧关键点信息,那在视频中的检测速度会有较大幅度的提高. 最新的一项工作,3DMaskR-CNN[14],在对每帧关键点检测的同时,也会通过 3D 卷积利用时序信息产生姿态轨迹流进行关键点预测,但是 3D 卷积融合帧间信息的同时,也带来了较大的计算量.

本文提出了一种在视频中进行人体关键点检测的方法,方法分为两部分,第一部分是对单帧图片进行人体关键点检测,对于上采样不能弥补空间分辨率损失以及常用检测模型速度较慢的情况,采用双路卷积神经网络,一路采用金字塔网络结构保持语义表达能力,并选轻量级模块 Inverted residuals 模块作为网络基础模块,提高模型运行速率,另一路保持空间分辨率以减少由于分辨率变化导致的损失. 第二部分,利用时间序列关系建立关键点轻量级跟踪模型,旨在利用视频中帧间信息进行视频人体关键点推理. 实验结果表明,本文的检测模型对比 Girdhar et al[15]、Xiu et al[16] 的 mAP 分别提高 16.1% 和 11%,对于图片单人检测速度分别提升 60% 和 50%,对于跟踪模块 MOT 分别有 16% 和 7.6% 提升. 本文对比最新的 Xiao B et al[9] 检测模型的 mAP,单人检测速度以及 MOT 分别有 1.3%,20% 以及 2.8% 的提高. 因此,本文的方法不仅可以增加在视频中行为姿态检测速率,同时还有效的提高了准确率.

2 轻量级双路卷积神经网络人体关键点检测

2.1 行人检测器

本文采用基于 SSD[17] 的物体检测算法. 为了减少参数并提高速率,采用高性能轻量级网络 MobilenetV2[18] 取代 SSD[17] 中的 VGG16[19] 作为骨架网络,并采用激活函数 ReLu6 如公式(1)取代 ReLu,提取主要的非线性特征.

$$f(x) = \min(6, \max(0, x)) \tag{1}$$

x 是激活函数的输入,$f(x)$ 是该函数的输出. ReLu6 是根据 ReLu 进行改进,将 ReLu 的最大范围控制在 6 以下,它常常使用在移动端设备中,主要是为了在移动端设备 float16 低精度的情况下,也会有很好的数值分辨率,如果对 ReLu 的激活范围不加限制,输出范围 0 到正无穷,如果激活值非常大,则输出分布会在一个很大的范围,此时在移动端设备中低精度的 float16 会无法很好地精确描述如此大范围的数值,从而带来精度损失.

为了训练物体探测器,在训练过程中仅使用 COCO[11] 数据集中的所有 80 个类别,选择所有类别是人类检测框作为本文多人人体关键点检测任务的输入.

2.2 关键点检测网络

作为一种基于神经网络的人体关键点检测算法,Xiao B et al[9] 在 2018 年 PoseTrack[20] 多人姿态估计挑战赛中取得了挑战赛的第二名的好成绩. 该方法主要是以 ResNet[21] 为主干网络,并根据网络输出的低分辨率特征层采用少量的转置卷积层来生成高分辨率特征层,最后构成了人体姿态估计中常见的沙漏结构. 姿态估计的输入是行人检测的结果,并将检测出的行人送入到人体关键点检测网络中,旨在检测 K 个关键点的位置,现阶段效果比较好的方法都是将回归出 K 个关键点的坐标转变为估计 K 个关键点的 heatmap,详见公式(2),输出关键点的 heatmap 的 channel 数为 K,即每一个 heatmap 表示第 K 个关键点位置的置信度.

$$f(x) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{2}$$

Xiao B et al[9] 采用的主干网络是 ResNet[21],ResNet[21] 是常用的图像特征提取的骨干网络之一,它采用 shotcut 方式缓解了深层网络中出现的梯度弥散和梯度爆炸的情况,它也常常用于姿态估计,比如在文献[15,16]中. ResNet[21] 因为其结构的复杂性和模块的多样性通常会导致更多的计算,这样会导致计算速度变慢.

MobilenetV2[18] 是一种新的基于移动端可以提高多项任务和标准测试的最新水平的神经网络. 该架构通过基于 Inverted residuals 结构,类似于 ResNet[21] 的 Residual block,采用深度卷积和点卷积代替普通卷积来提取特征,有效降低了模型参数和计算复杂度,同时实验结果得到相似的精度. 这种配置方法在理论上可以减少卷积层的时间复杂度和空间复杂度. 一个标准卷积层输入为 $D_k \times D_k \times M$ 的特征图 F,并得到一个 $D_f \times D_f \times N$ 的输出特征图 G,其中 D_f 表示输入特征图的宽和高,M 是输入的通道数(输入的深度) G 为输出特征图的宽和高,N 是输出的通道数(输出的深度). 标准卷积层计算量为 $D_k \times D_k \times M \times$

$N \times D_f \times D_f$,深度可分离卷积的计算量为: $D_k \times D_k \times M \times D_f \times D_f$,点卷积计算量: $M \times N \times D_f \times D_f$,则深度可分离卷积的计算量为 $D_k \times D_k \times M \times D_f \times D_f + M \times N \times D_f \times D_f$. 通过将普通卷积分为滤波和组合的过程得到对计算量的缩减,缩减比例详见公式(3),对计算速度有较大幅度提升.

$$\frac{D_k \times D_k \times M \times D_f \times D_f + M \times N \times D_f \times D_f}{D_k \times D_k \times M \times N \times D_f \times D_f} = \frac{1}{N} + \frac{1}{D_k^2} \quad (3)$$

MobileNetV1 使用深度可分离卷积作为网络基本模块,但在实用中会出现特征退化和梯度消失,MobilenetV2 [18] 主要学习了 ResNet [21] 的思想并基于深度可分离卷积采用 shotcut 模式,防止梯度消失,同时采用一个 1×1 卷积核提升通道数,以防止特征退化,在低纬度空间,线性映射会保存特征,而非线性映射会破坏特征,所以使用 linear 激活函数代替 ReLu 激活函数来增加信息保留,具体结构见图 1. 因此根据以上优点,本文选取 MobilenetV2 [18] 作为 backbone.

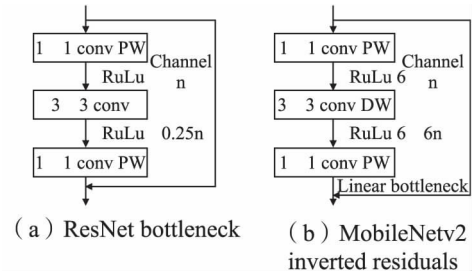

图 1　ResNet 结构单元与 MobileNetv2 结构单元对比图

Fig. 1　Comparison between ResNet structural unit and MobileNetv2 structural unit

Xiao B et al [9] 主要是高分辨率特征图下采样至低分辨率,再从低分辨率特征图采用转置卷积层来生成高分辨率的思路,但是在这类网络中,高分辨表征主要是低分辨率表征通过上采样得到的高分辨率表征,其本身虽然拥有很好的语义表达能力,但是上采样本身并不能完整地弥补空间分辨率的损失. 所以,最终输出的人体姿态估计高分辨率表征所具有的空间敏感度并不高,空间敏感度很大程度上受限于语义表达力强的表征所对应的分辨率. 所以本文希望在整个网络过程中始终保持高分辨率表征,同时增加高分辨率和低分辨率之间进行信息交换和融合,从而希望可以得到足够的丰富语义信息.

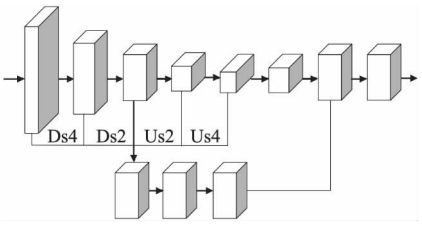

图 2　人体关键点检测网络结构图

Fig. 2　Network structure diagram of Human keypointdetection

本文采用了两路分支,第一路分支主要以 MobilenetV2 [18] 为主干网络,下采样 32 倍. 同时构建金字塔网络,选用下采样 8 倍特征图大小为基准,将下采样 2 倍,4 倍特征图进行 3×3 卷积变成下采样为 8 倍特征图,同时通过转置卷积对特征图进行上采样为 8 倍特征图,然后将各个部分的特征图进行拼接融合作为第二路分支的输入. 之后第二路经过卷积处理后和第一路的特征图进行拼接融合,最后添加 1×1 卷积核用来生成所有 k 个身体关键点的预测热力图 $\{H_1 \cdots H_k\}$. 具体结构见图 2, 'Ds' 表示下采样, 'Us' 表示上采样.

$$l2loss = \frac{1}{n \times m} \sum_{m}^{n} \sum_{j}^{m} | f(x_{ij}) - Y_{ij} |^2 \quad (4)$$

与文献 [21] 中相同,采用均方误差(MSE)作为损失函数,详见公式(4),$f(X_{ij})$ 为预测热力图的概率值,Y_{ij} 为热力图的标签值,图像在第 k 个关键点周围生成高斯分布作为热力图标签,与网络预测的热力图概率图进行比较,测试时选择在热力图上概率最大点的坐标作为该关键点的坐标值.

2.3　关键点帧间推理网络

本文提出一个新颖的使用帧间信息进行人体关键点跟踪方法(如图 3 所示),基于先前帧的关键点的位置信息和关联信息使用卷积神经网络的方法建立信道连接,最后对预测帧的关键点信息进行推理.

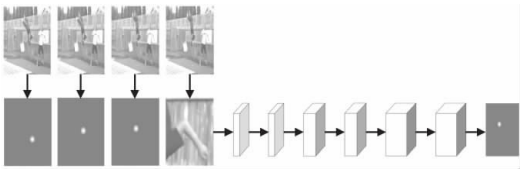

图 3　人体关键点跟踪网络结构图

Fig. 3　Network structure diagram of Human keypointtracking

步骤 1. 本文使用人体检测器对图像进行行人检测,从而得到每张图中行人的位置.

步骤 2. 将用 2.2 介绍的模型作为人体关键点检测模型,并将步骤 1 得到的行人提取出来,作为人体关键点模型的输入.

步骤 3. 根据人体关键点检测算法记录前 n 帧每帧中人物中关键点的位置 (x_i, y_i),并记录下每个 j 关键点在这几帧中的坐标,然后根据坐标得到时序坐标中的最值,即 $(x^j_{\min}, y^j_{\min}, x^j_{\max}, y^j_{\max})$,具体见公式(5)-公式(8).

并根据时序坐标从预测帧中抠出局部图片,并对局部图片的关键点生成该点热力图,并与待测图片的相同范围内建立时序信道,作为跟踪模型的输入. 为了增大局部感受野的范围,将时序信道的范围扩大 1.3 倍.

$$x^j_{\min} = \min\left(\sum_{1}^{n} x^j_i\right) \quad (5)$$

$$y^j_{\min} = \min\left(\sum_{1}^{n} y^j_i\right) \quad (6)$$

$$x^j_{\max} = \max\left(\sum_{1}^{n} x^j_i\right) \quad (7)$$

$$y^j_{\max} = \max\left(\sum_{1}^{n} y^j_i\right) \quad (8)$$

步骤 4. 网络模型是基于 MobilenetV2 [18] 的 Inverted residuals 结构的 6 层卷积网络. 第一层通过 Inverted residuals 获得 8 个特征映射通道,然后通过每两个 Inverted residuals,将得到的特征图通道扩大原来的两倍,经过 6 个 Inverted residuals,特征通道扩大为第一层特征层的 32 倍,之后经过一个 1×1 的卷积进行维度降维,获得具有 1 个通道的预测热力图,表 1 为关键点帧间推理网络结构参数,采用 l2 损失对预测帧进行

比较. 图 3 提供了设计中跟踪器网络的示意图.

由于在人体关键点检测中发现分辨率大小的改变在增加感受野的同时,也会带来精度方面的损失,而且在跟踪关键点时,对于关键点来说,主要在意的是关键点周围的信息,而不是全局信息,所以本文中关键点跟踪网络主要是分辨率保持一致. 这样既保持周围信息的同时,减少由于分辨图大小不同带来 benchmark 的差异.

表 1 关键点帧间推理网络结构

Table 1 Key point inter frame reasoning network structure

卷积名	Con1	Con2	Con3	Con4	Con5	Con6	Con7
卷积核	3×3	3×3	3×3	3×3	3×3	3×3	3×3
通道数	8	8	16	16	32	32	1

步骤 5. 根据步骤 3 可以得到一个关键点预测,如果想要知道其它关键点的预测,根据步骤 3 重新建立时序信道,并采用步骤 4 跟踪模型跟踪关键点.

由于人体关键点检测需要较高的模型复杂度才可以得到精确的结果,通常需要很高的计算量并需要花费较多的时间,建立跟踪模型旨在利用已知的每帧信息和帧间信息对预测帧进行信息推理,减少单张图片中人体关键点检测所要花费的时间. 本文实验根据前 3 帧关键点的热力图和预测帧进行融合,建立时序信道作为网络的输入,同时采用 6 层 MobilenetV2 [18] 的 Inverted residuals 以及 1 个点卷积作为本文的网络骨架,在精度保持的同时减少计算量. 在 PoseTrack 数据集 [20] 中,本文为每个行人建立了 12 个关键点跟踪器,并在测试集中进行验证,实现良好的效率,这将在第下一章中进行分析.

3 实验结果与分析

3.1 实验数据集介绍

本文采用的是 PoseTrack 数据集 [20],该数据集是用于大规模多人姿势估计和跟踪野外视频的基准. 它共包含 550 个视频,共 66,374 个帧,分别包含 292 个训练集,50 个验证集和 208 个测试集. 训练集中视频的中间 30 帧密集地标有人体关键点. 对于验证和测试视频,除了中间 30 帧之外,每四帧也被注释用于评估远程关节跟踪. 总的来说,数据集包含 23,000 个标记帧和 153,615 个姿势.

数据集旨在评估三种不同任务的方法. 任务 1 使用平均精度(mAP)度量来评估单帧姿态估计,如文献 [22] 中所做的那样. 任务 2 还评估视频中姿势估计. 任务 3 使用多对象跟踪度量(MOT)如文献 [23] 评估跟踪. mAP 的计算方法如文献 [2] 所示,MOT 如文献 [24] 中所述. 本文使用 PoseTrack 数据集 [20] 评估系统计算论文中提供的所有结果.

3.2 实验环境

实验服务器操作系统为 Ubuntu 16.04 Server,配置有 AMD 锐龙 2700 3.2 GHz CPU,和 12G 显存的华硕 GTX1080ti 显卡. 行人检测、人体关键点检测以及人体关键点跟踪都是使用了开源的深度学习框架 tensorflow.

3.3 实验设置

由于人体关键点检测的输入是行人检测的输出,本文需要从得到的图像中裁剪并调整为固定分辨率,默认分辨率为 256:192. 为了获得更大的人体关键点感受野,将行人检测的边界框延长 15%(两侧为 7.5%)来获得行人的位置. 数据增强包括旋转(±40 度),调整比例(±30%)和翻转等. 本文的 MobileNetV2 [18] 主干网络通过在 COCO [11] 数据集进行了预训练,训练分为 140 个 epoch,起始学习率为 e^{-3},学习率在 90 个 epoch 的时候下降到 e^{-4},在 120 个 epoch 下降到 e^{-5}. 用 PoseTrack 数据集 [20] 训练检测模型时,先用 COCO 数据集的训练结果作为预训练模型再进行 finetune,训练总计分为 20 个 epoch,开始学习率为 e^{-3},在 10 个 epoch 下降到 e^{-4},在 15 个 epoch 下降到 e^{-5}. 同时,Mini-batch 大小为 64,使用 Adam 优化器.

训练跟踪模型时,本文主要建立了 12 个关键点的跟踪模型,分别为左肩,右肩,左肘,右肘,左手腕,右手腕,左髋,右髋,左膝,右膝,左脚踝和右脚踝. 对 PoseTrack 数据集 [20] 时序帧进行切分,根据每四张图片时序帧中前三已知帧的关键点的位置来切分第四帧的坐标. 在训练时,本文增加了与关键点模型相同的数据增强,例如旋转(±40 度),比例(±30%)和翻转. 总共有 100 个 epoch,基础学习率为 e^{-3},它在 50 个 epoch 下降到 e^{-4},在 80 个 epoch 下降到 e^{-5}. 本文的模型损失函数是 L2loss.

3.4 实验分析

轻量级双路卷积神经网络的人体关键点检测方法和时间序列帧间推理网络的人体关键点跟踪方法,为说明模型的有效性,本文对于人体关键点检测从 mAP 和运行时间上进行评估.

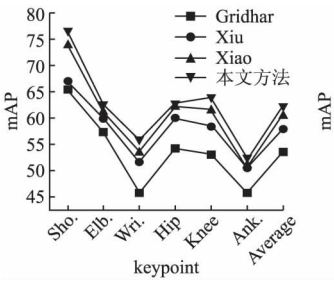

图 4 不同算法在 val mAP 比较

Fig. 4 val mAP comparison of different algorithms

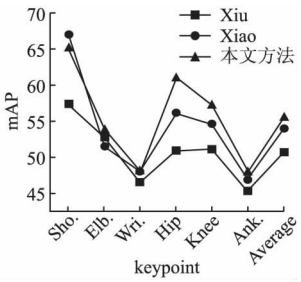

图 5 不同算法在 test mAP 比较

Fig. 5 test mAP comparison of different algorithms

图 4 和图 5 报告了姿态估计的结果(任务 2). 本文的人体关键点检测网络除了脚踝关键点,其他关键点的 mAP 相对其它的方法有较大的改善,在验证集中,本文方法分别比 Girdhar et al [15],Xiu et al [16] 的 mAP 提高了 11.2 和 4.8,约 21.1% 和 7.2%. 对比 Xiao B et al [3] 的 mAP 提高了 0.4. 在测试集,本文方法比 Xiu et al [16] 的 mAP 提高了 6.8,约 11%. 对比 Xiao B et al [3] 的 mAP 提高了 0.9,依据图 4 和图 5 可知:在测试集和验证集中,本文的算法对于大多数关键点都会有较大幅度的提升,对于踝关节关键点会有略微下降,主要是由于左右踝关节相似度比较高导致学习难度系数比较高.

表 2 不同算法单人图像时间对比

Table 2 Time comparison of single person with different algorithms

模型	Girdhar et al [15]	Xiu et al [16]	Xiao B et al [3]	本文方法
时间/s	0.32	0.30	0.25	0.20

表 2 报告了对于单人图像姿态估计的运行时间. 本文的

模型相对于其他方法有了较大幅度的提高，对 Gridhar[15] 和 Xiu[16] 平均分别减少 0.12s 和 0.10s，前向速度分别提升 60% 和 50%，对于 Xiao B[9] 算法平均减少了 0.05s，前向速度提升 25%. 总的来说，本文算法相比其它最近的工作有了较大的改进.

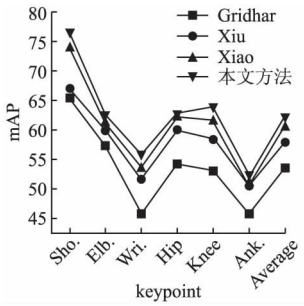

图 6　不同算法在 val MOTO 比较

Fig. 6　Moto comparison of different algorithms in val

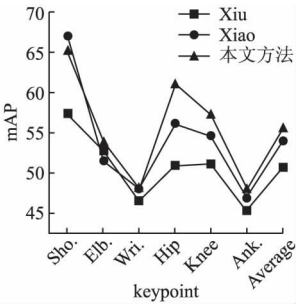

图 7　不同算法在 test MOTO 比较

Fig. 7　Moto comparison of different algorithms in test

在验证集中，本文的模型在 MOTO 相比 Gridhar[15]，Xiu[16] 分别提高 3.7 和 4.8，约 7.1% 和 9.5%，对比 Xiao B[9] 的 MOTO 提高了 1.5，约 2.8%. 在测试集中，本文的模型在 MOTO 相比 Xiu[16] 提高 4.4，约 7.6%，对比 Xiao B[9] 的 MOTO 提高了 1.7，约 2.8%. 总的来说，本文的模型总体上优于其它方法，如图 6 和图 7 所示.

表 3　不同网络骨架在 PoseTrack 测试集性能比较

Table 3　Performance comparison of different network backbone in PoseTrack Test

模型	6 层 Depth-wise Separable Convolution	6 层 Residual block	6 层 Inverted residuals
时间/s	0.016	0.02	0.017
MOTO	48.3	56.7	55.5

表 3 报告了常用骨架下跟踪模型的性能. 本实验主要是使用常用的人体关键点检测骨架 ResNet[21] 的 Residual block、轻量级网络 MobileNetv1 的深度卷积 + 点卷积模块，简称 Depth-wise Separable Convolution，和本文选择使用的 6 层 Inverted residuals 模块进行对比，由实验结果来看，选择同样层数的结构，ResNet[21] 由于由于本身参数较多，跟踪一个关键点的时间大约是 0.02s，大约是本文使用的 Inverted residuals 的 1.2 倍，与之同时，它的 MOTO 也是最高的与本文算法的 MOTO 基本持平. 得出结论本文提出的骨架比 ResNet[21] 在精度基本持平的同时运行时间更短. 同时，本文选择了同样是轻量级网络 MobileNetv1 Depth-wise Separable Convolution 模块进行比较，速度提升大约 6.25%，实验看出这套组合比 Inverted residuals 模块时间更快，提升了 0.001s，但是精度相比来说下降 14.9%，由于 Depth-wise Separable Convolution 是 Inverted residuals 的基础模块，计算量和参数都比较小，但是由于结构较为简单并没有很好的融合高维特征，精度并没有达到 Inverted residuals 相同水平. 总的来说 Inverted residuals 相比 Depth-wise Separable 精度和时间结合更有效. 通过实验结果表明本文提出的算法在性能指标上都取得了比主流算法要好的结果.

4　结　论

本文主要探究了针对视频中姿态估计缺少高分辨率表征和时效性的特点，提出了基于传统的神经网络姿态估计“下采样上采样”模型，提出了一种轻量级双路金字塔卷积神经网络来提高姿态估计的高分辨率表征，并利用帧间信息推理网络，并选用轻量级模块预测关键点信息来提高时效性. 本文方法主要在 PoseTrack 数据集[20] 上进行训练和验证. 精度和速度上相比最近 Gridhar[15]，Xiu[16]，Xiao B et al[9] 较大提高，证明本文的思路方法真实有效.

本文帧间信息推理实验主要采用了 Depth-wise Separable Convolution、Residual block、Inverted residuals 模块，但是对于如何更好的利用时序信息还有提高的空间. 与此同时，本文发现采用前三帧的信息预测效果较好，但如何利用更少的前帧信息预测关键点信息也是以后工作重点提高的地方. 同时由于本文采用的是每次预测局部单个关键点信息，在后继研究中，将探索利用图片全局信息，增加帧与帧之间不同关键点的相互作用. 进一步优化网络模型并提高在视频中的检测性能.

References:

[1] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition [C] //Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

[2] Yang W, Li S, Ouyang W, et al. Learning feature pyramids for human pose estimation [C] //IEEE International Conference on Computer Vision (ICCV), 2017: 1281-1290.

[3] Xiao T, Li H S, Ouyang W, et al, Learning deep feature representations With domain guided dropout for person re-identification [C] //IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 1249-1258.

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

[5] Chen Yao-dong, Liu Qin, Peng Die-fei. Component-aware adaptive algorithm for pose estimation [J]. Computer Engineering, 2018, 44 (11): 257-264.

[6] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN [C] //IEEE International Conference on Computer Vision (ICCV), 2017: 2961-2969.

[7] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks [C] //IEEE Conference on Computer Vision and Pattern Recognition (ICCV), 2014: 1653-1660.

[8] Wei S H, Ramakrishna V, Kanade, et al. Convolutional pose machines [C] //IEEE Conference on Computer Vision and Pattern Recognition (ICCV), 2016: 4724-4732.

[9] Xiao B, Wu H P, Wei Y C. Simple baselines for human pose estimation and tracking [C] //European Conference on Computer Vision ((ECCV)), 2018: 466-481.

[10] Chen Y L, Wang Z C, Peng Y X, et al. Cascaded pyramid network for multi-person pose estimation [C] //IEEE Conference on Computer Vision and Pattern Recognition (ICCV), 2018: 7103-7112.

[11] Lin T, Maire, Belongie S, et al. Microsoft COCO: common objects in context [C] //European Conference on Computer Vision (EC-

CV) ,2016: 740-755.

[12] Cao Z,Simon T,Wei S,et al. Realtime multi-person 2D pose estimation using part affinity fields [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2017: 7291-7299.

[13] Lin T Y,Dollar P,Girshick R,et al. Feature pyramid networks for object detection [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2017: 2117-2125.

[14] Chu X,Yang W,Ouyang W L,et al. Multi-context attention for human pose estimation [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2017: 1831-1840.

[15] Girdhar R,Gkioxari G,Torresani L,et al. Detect-and-track: efficient pose estimation in videos [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2018: 350-359.

[16] Papandreou G,Zhu T,Kanazawa N,et al. Towards accurate multi-person pose estimation in the wild [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2017: 4903-4911.

[17] Liu W,Anguelov D,Erhan D,et al. SSD: single shot multibox detector [C] //European Conference on Computer Vision(ECCV) , 2016: 21-37.

[18] Sandler M,Howard A,Zhu M L,et al. MobileNetV2: inverted residuals and linear bottlenecks [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2018: 4510-4520.

[19] Karen S,Andrew Z. Very deep convolutional networks for large-scale image recognition [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2015.

[20] Andriluka M,Iqbal U,Insafutdinov E,et al. PoseTrack: a benchmark for human pose estimation and tracking [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2018: 5167-5176.

[21] He K M,Zhang X Y,Ren S Q,et al. Deep residual learning for image recognition [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2017: 770-778.

[22] Chu X,Ouyang W,Li H S,et al. Structured feature learning for pose estimation [C] //IEEE Conference on Computer Vision and Pattern Recognition(ICCV) ,2016: 4715-4723.

[23] Liu Z,Zhu J K,Bu J J,et al. A survey of human pose estimation: The body parts parsing based methods [J] //Journal of Visual Communication and Image Representation,2015,32: 10-19.

[24] Chen Y,Shen C H,Wei X S,et al. Adversarial posenet: a structure aware convolutional network for human pose estimation [C] //IEEE International Conference on Computer Vision(ICCV) ,2017: 1212-1221.

附中文参考文献:

[5] 陈耀东,刘 琴,彭蝶飞. 面向姿态估计的组件感知自适应算法 [J] 计算机工程,2018,44(11) : 257-264.