文章编号：2095-2163(2020)07-0137-05 中图分类号：TP391 文献标志码：A

基于回归关节点偏移量与热力图的多人姿态估计算法

范冬艳，王 倩
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要： 人体姿态估计大多采用回归关节点坐标或分类预测关节点热力图的方式，容易造成关节点误判及精度受限；对于多人解析，传统方法是从所有的候选关节点出发，进行关节点匹配，容易导致错误的连接。本文提出了一种基于回归关节点偏移量与热力图的多人姿态估计算法，算法主要由两个阶段组成：第一阶段 encoder 层，对于关节点检测，同时分类预测关节点的热力图和回归坐标 2-D 偏移向量，精确定位关节点位置，对于关节点关联，使用部件关联字段，具有在低分辨率激活图上存储细粒度信息的能力；第二阶段 decoder 层，利用 Hopcroft-Carp 算法对进行人体姿态解析，将一个 K 分图匹配问题转变为二分图匹配，能够极大提高准确率且减少时间复杂度。

关键词： 人体姿态估计；部件关联字段；二分图匹配

Multi-person Attitude Estimation Algorithm Based on Regression Joint Offset and Thermal Diagram

FAN Dongyan, WANG Qian
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 In order to solve the problem of joint misjudgment and limited accuracy, joint coordinates are mostly regressed or classified to predict joint thermal diagram. For multi-person analysis, the traditional method matches joint points from all candidate joint points, which easily leads to the wrong connection. A multi-person attitude estimation algorithm based on regression joint offset and thermal diagram is proposed. The algorithm is mainly composed of the following two stages: the first stage is composed of encoder layers. For joint point detection, it classifies and predicts the joint point thermal diagram and regression coordinate 2-D offset vector at the same time and accurately locates the joint point position. For joint point association, it uses the component association field, and has the ability to store fine-grained information on the low-resolution activation map. In the decoder layer of the second stage, Hopcroft-Carp algorithm is used to analyze the human posture. Hopcroft-Carp algorithm transforms a K-map matching problem into a bipartite matching, which can greatly improve the accuracy and reduce the time complexity.

【Key words】 Human pose estimation; 2-D offset; Hopcroft-Carp algorithm

0 引 言

人体姿态估计是计算机视觉的一个重要研究领域，指在一张拥有多人的 RGB 图像中准确定位人体关节点的像素位置[1]。相对单人姿态估计，多人姿态估计存在更大的挑战性，因为图像和视频中的交互、遮挡等问题更为突出。多人姿态估计根据预测过程分为 Top-down 和 Bottom-up。Top-down 先检测出单个人的边界框，再分别检测每个人的关节点[2]；Bottom-up 先进行人体关节点检测，再利用关联算法对关节点进行关联[3]。

考虑到 Top-down 受检测框的影响较大，易产生漏检误检，且处理速度慢，实时性不强，本文采取 Bottom-up 的检测方式。对于 Bottom-up 的关节点检测方法，大多采用回归关节点坐标或分类预测关节点热力图的方式，容易造成关节点误判及精度受限；对于多人解析，传统方法是从所有的候选关节点出发，进行关节点匹配，容易导致错误的连接问题。

针对 Bottom-up 存在的问题，本文提出了一种基于回归关节点偏移量与热力图的多人姿态估计算法。算法主要由两个阶段组成：第一阶段 encoder 层，对于关节点检测，同时分类预测关节点的热力图和回归坐标 2-D 偏移向量，精确定位关节点位置，对于关节点关联，使用部件关联字段，具有在低分辨率激活图上存储细粒度信息的能力；第二阶段 decoder 层，利用 Hopcroft-Carp 算法对进行人体姿态解析，将一个 K 分图匹配问题转变为二分图匹

基金项目： 国家自然科学基金青年基金项目(61802251)；上海市科学技术委员会科研计划项目(16dz1206000)；上海工程技术大学科研项目(E3-0501-18-01043)。

作者简介： 范冬艳(1995-)，女，硕士研究生，主要研究方向：计算机视觉与图形处理；王 倩(1995-)，女，硕士研究生，主要研究方向：计算机视觉与图形处理。

收稿日期： 2020-03-24

配,能够极大提高准确率,且减少时间复杂度。

1 网络结构概述

本文提出的算法模型如图 1 所示。模型输入为一系列经过预处理的图像数据,模型主干由二部分组成:第一部分是编码层(Encoder),第二部分是解码层(Decoder)。Encoder 主要通过 7×7 Conv、2×2 max pooling 和 Densenet 进行特征提取,生成热力图(heatmap)、2-D 偏移向量(offset)和部件关联字段(PAF)。其中,关节点检测器由热力图和 2-D 偏移向量融合而成;关节点关联器由 PAF 生成;Decoder 通过 Hopcroft-Carp 算法进行多人解析,以生成完成的人体姿态。

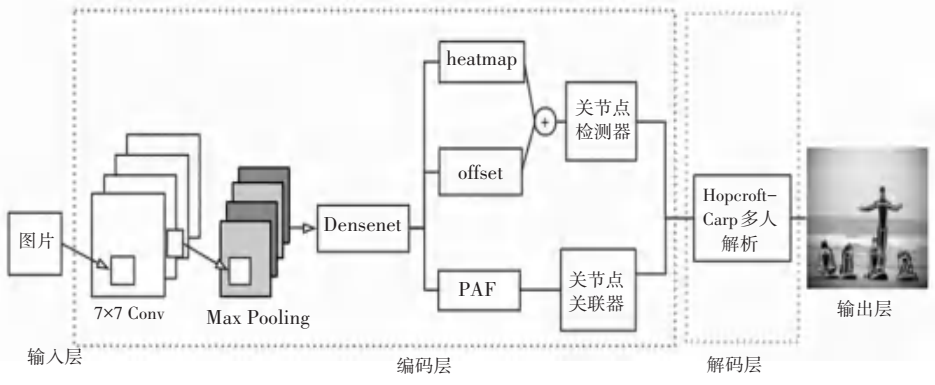


图 1 模型架构

Fig. 1 Model Architecture

1.1 编码层

1.1.1 关节点检测器

对于 Bottom-up 的人体关节点检测,常用方法有两种:

(1)通过回归得到关节点的坐标,此方法在处理多人姿态估计时,若图像中人数大于 1,容易引起关节点误判;

(2)分类预测关节点坐标的高斯响应热力图。对图片中每个像素位置赋予一个概率,像素点与关节点距离越近,其概率越接近于 1,高斯热力图响应越大。这种方法虽然解决了第一种方法关节误判的问题,但其定位精度受网络输出特征图大小的限制。

为解决以上两种方法的局限性,本文提出了回归和分类相结合的关节检测方法。关节点检测器由二分支组成,第一分支预测关节点热力图,即对每一个空间像素点进行分类,判断其是否在关节点附近,热力图的通道数为 k , k 为关节点个数。第二分支预测关节点 2-D 偏移向量(x-offset,y-offset),以便精确估计关节点位置。因为每个关节点都有两个方向的偏移通道,所以总通道数为 $2k$ 。从位置 i 到关节点 k 处的偏移量如公式(1)和公式(2)所示:

x - offset:

$$F_k(x_i) = l_k - x_i \tag{1}$$

y - offset:

$$F_k(y_i) = l_k - y_i. \tag{2}$$

融合方式如图 2 所示,分别提取左肘关节的热力图与朝关节点标签处的偏移向量场,对其进行融合,以产生高度局部化的热力图。

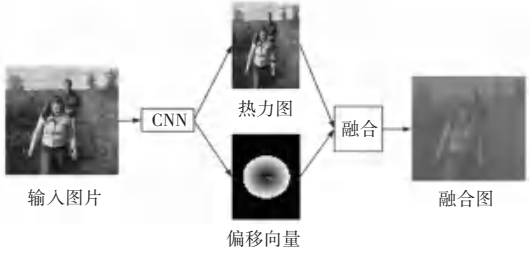


图 2 热力图与偏移量融合图

Fig. 2 Heatmap and offset fusion diagram

在生成关节点热力图和 2-D 偏移向量后,需对二者进行融合,以生产精确的坐标估计,具体融合方式如公式(3)所示:

$$f_k(x_i) = \sum_j \frac{1}{\pi R^2} G(x_j + F_k(x_j) - x_i) h_k(x_j). \tag{3}$$

其中, $G(.)$ 是双线性插值核, R 表示 x_i 距 l_k 在半径为 R 的圆环内。

1.1.2 关节点关联器

在人群拥挤的场景中,将关节联合成多种姿势极具挑战性。本文采用部件关联字段(Part Association Fields,PAF)方法进行关节点关联,此方法具有在低分辨率激活图上存储细粒度信息的能力。

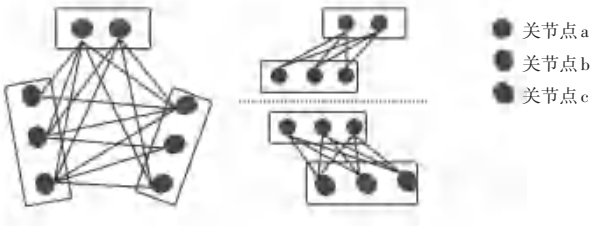
PAF 是一种复合结构,其输出为 $(b,h,w,k,7)$, k 为关键点个数, (h,w) 为输出特征图的大小,7 代表 7 个元素: $\{a_c^{ij}, a_{x1}^{ij}, a_{y1}^{ij}, a_{b1}^{ij}, a_{x2}^{ij}, a_{y2}^{ij}, a_{b2}^{ij}\}$,其中, a_c^{ij}

表示位置(i,j)的置信度，$(a_{x1}^{ij},a_{y1}^{ij})$，$(a_{x2}^{ij},a_{y2}^{ij})$分别表示该像素点到标签的偏移向量，$a_{b1}^{ij},a_{b2}^{ij}$用于计算损失。

PAF 通过两步进行关节点关联。首先，找到离任一像素点(i,j)最近的关节点；其次，根据该关节点的标签，找到与其相连的另一关节点。

1.2　解码层

在编码层得到关节点位置和关节点关联后，本文在解码层通过 Hopcroft-Carp 算法进行多人解析，把同属于同一个人的线段都连起来。通常，传统的方法是从所有的候选关节点出发，进行最优关节点匹配，把此问题转变为一个求解 K 分图的最优解问题，会导致属于同一个人的关节点会形成一个错误的连接问题[4]。因此，本文选择对相邻的两个关节点之间匹配，例如手肘和手腕之间的匹配，这样一个 K 分图匹配问题便转化为二分图匹配问题。如图 3 所示，假设关节点 a 与关节点 b 具有相连性；关节点 b 与关节点 c 具有相连性；关节点 a 与关节点 c 没有相连性，若采用图 3(a) 的 K 分图匹配，会导致 a 与 c 相连；若采用如图 3(b) 的二分图匹配，则不会出现该问题。



(a) K 分图　　(b) 二分图

图 3　K 分图与二分图匹配对比

Fig. 3　Comparison between K-Score and bipartite matching

二分图匹配定义：给定一个二分图 G，若边集 G 的一个子集 M 满足 G 中任意两条边 E 都没有相同的顶点 V，则称 M 是一个匹配。人体估计的二分图 G，如图 3 所示，用 X,Y 分别表示两个不同关节点位置的集合，x_i,y_i 分别表示 X 和 Y 中的第 i 个关节点。

曹哲采用匈牙利算法解决二分图匹配问题[5]，但是为了降低时间复杂度，可以在增广匹配 M 时，多次寻找多条增广路径，因此本文选用 Hopcroft-Carp 算法，算法流程如下：

(1) 从图 $G=(X,Y)$ 中，任意取一个 M 作为初始匹配；

(2) 若 M 包含 X 中的所有顶点，返回 M；否则，以所有未匹配的顶点为源点，进行一次广度优先搜索(BFS)，并标记每个顶点到源点的距离；

(3) 若边集 $<v,u>$ 满足 $dis[v]=dis[u]+1$，则从 X 中找一个未被 M 匹配的顶点 x_0，记 $S:\{x_0\}=\varnothing$；

(4) 若 $N(S)=T$，返回；否则取 $y_0\in N(S)-T$；

(5) 若 y_0 已被 M 匹配，转步骤(6)，否则做一条 M - 增广路径 $P(x_0,y_0)$，取 $M=M\Delta P(x_0,y_0)$；

(6) 因为 Y 已被 M 匹配，所以 M 存在一条边 (y_0,z_0)，令 $S=S\cup\{z_0\}$，$T=T\cup\{y_0\}$，转步骤(2)。

最终，在 Hopcroft-Carp 算法进行人体关节点匹配的基础上，按照人体的结构化信息对连接匹配度高的候选肢体完成多人姿态估计。该算法在寻找增广路径的同时寻找多条不相交的增广路径，形成极大增广路径集，然后对极大增广路径集进行增广。在寻找增广路径集的每个阶段，找到的增广路径集都具有相同的长度，且随着算法的进行，增广路径的长度不断扩大。可以证明，最多增广$\sqrt{n}$次(已做修改) 就可以得到最大匹配。

2　实　验

2.1　数据集与评价标准

2.1.1　MS COCO 数据集

MS COCO 全称 Microsoft Common Objects in Context，下文简称 COCO，专门用于对象检测和分割、语义分割、字幕生成、人体关节点检测等任务。COCO 数据集共有 33 万张图像，其中拥有关节点标注的人共有 25 万个。此外，为方便训练和测试，COCO 数据集提供了 Python、Matlab 和 Lua 的 API 接口，可以提供完整的图像标签数据的加载，解析和可视化。数据集对于每个人体的关节点标注共有 17 个，分别为：鼻子、左眼、右眼、左耳、右耳、左肩膀、右肩膀、左手肘、右手肘、左手腕、右手腕、左髋部、右髋部、左膝盖、右膝盖、左脚踝、右脚踝。

2.1.2　评价标准

COCO 关键点检测的评价指标类似于对象检测，即平均精确率(Average Precision AP) 和平均召回率(Average Recall，AR) 及其变体，AP 表示正确识别物体的个数占总识别物体个数的百分比，AR 表示正确识别物体的个数占测试集中物体的总个数的百分比。

(1) AP 与 AR 计算方式。真实标签中含有正例和反例，设正例中预测结果为正例的个数为 TP；正例中预测的结果为反例的个数为 FN；反例中预测结果为正例的个数为 FP；反例中预测结果为反例的个数为 TN。则 AP 和 AR 的计算方法分别为公式(4) 和公式(5)。

$$Precision = \frac{tp}{tp + fp}, \quad (4)$$

$$Recall = \frac{tp}{tp + fn}. \quad (5)$$

mAP(mean Average Precision) 和 mAR(mean Average recall)即是针对于所有类别的 AP,对应到本文 COCO 的关节点估计,AP 和 AR 针对是对单个关节(如手腕)做平均,而 mAP 表示对 COCO 数据集 17 个关节点的 AP 和 AR 再做平均。

(2)OKS。AP 和 AR 的核心是真实标签与预测目标之间的相似性度量。为此,定义了一个对象关键点相似性(Object Keypoint Similarity ,OKS)。

每个对象的真实关节点格式为:$[x_1, y_1, v_1, \ldots, x_k, y_k, v_k]$,$x, y$ 分别表示关节点的横纵坐标,v 是可见标志,$v = 0$ 表示未标注点,$v = 1$ 表示有标记但图像中不可见(比如被遮挡),$v = 2$ 表示标记且图像中可见,实际预测时不要求预测每个关节点的可见性。每个对象的关键点检测器必须输出关键点位置和对象级别的置信度。对象的预测关键点具有与实际真值相同的形式:$[x_1, y_1, v_1, \ldots, x_k, y_k, v_k]$。 评估时,检测器的预测 v_i 并未使用,即不需要关键点检测器来预测每个关键点的可见性或置信度。OKS 定义为公式(6):

$$OKS = \frac{\sum_i \exp\{-d_{p^2}^2 / 2S_p^2\sigma_i^2\}\delta(v_{p^i} = 1)}{\sum_i \delta(v_{p^i} = 1)}. \quad (6)$$

其中,p 表示在标签中某个人的 id;p^i 表示某个人的关节点 id;v_{p^i} 表示这个关节点的可见性为 1;S_p^2 表示这个人所占的面积大小平方根,根据标签计算得到;σ_i 表示第 i 个关节点的归一化因子,这个因子是通过对已有的数据集中所有标签计算的标准差而得到的,反映出当前骨骼点对与整体的影响程度,值越大,说明在整个数据集中对这个点的标注效果越差;值越小,说明整个数据集中对这个点的标注效果越好。

2.2 数据集预处理

对于 COCO 数据集的增强,在训练期间使用介于 1.0 和 1.5 之间的重缩放因子。首先,本文使用 1.25 的缩放因子调整图像大小;其次,对重缩放的图像进行裁剪,以固定的其像素大小为(353,257),图像的纵横比值为 353/257=1.3;最后,调整图像相应边界框的坐标,在此期间随机对图像进行填充操作,填充大小是图像的 0.1~1.0 倍,填充和边界框的 IOU 值在[0.1,0.3,0.5,0.7,0.9]这几个值中随机选取。

2.3 实验环境与参数设置

(1)实验环境。本文选用 TensorFlow 作为训练框架,并配置 Ubuntu16.4 + python3.5 + anaconda3.0 + pycharm + vs2015 + cuda9.0 + cuDNN7.0 的训练环境。TensorFlow 是一款开源的深度学习框架,支持多种客户端语言下的安装和运行,具有很强的可视化功能和多个用于高级模型开发的选项,其内置的 Tensorboard 用于对网络进行可视化训练。

(2)参数设置。编码层将空间分辨率加倍的单层 1×1 Conv,使用 Sigmoid 置信度分量进行归一化。使用交叉熵损失对融合 offset+heatmap 过程进行训练,并使用 L1 损失用于 PAF 的训练。使用 SGD 优化器,批标准化为 24,学习率为 0.000 1,动量为 0.95,不设置衰减权重。

3 实验结果

(1)准确率与召回率对比分析。为了证明本文提出的深度学习的多人姿态估计算法能够有效地提高关节点识别与关节点关联性能。将本文算法与现有的目标检测算法进行对比,对比结果见表 1。

表 1 coco 数据集评价结果

Tab. 1 Coco dataset evaluation results

方法	AP	$AP_{0.50}$	$AP_{0.75}$	AR	$AR_{0.50}$	$AR_{0.75}$
Top-down						
MASK-RCNN	0.638	0.849	0.675	0.697	0.916	0.749
G-RMI	0.649	0.855	0.713	0.697	0.887	0.755
Bottom-up						
Personlab	0.687	0.890	0.754	0.754	0.927	0.812
PifPaf	0.787	0.950	0.887	0.864	0.960	0.854
Ours	**0.824**	**0.979**	**0.921**	0.893	0.971	0.872

本文提出的算法是一种 Bottom-up 的多人姿态估计算法,不仅选择 Personlab、PifPaf 作为对比算法,还选择 Top-down 的算法 MASK-RCNN 和 G-RMI 进行比对。MASK-RCNN 和 G-RMI 是近年来较为热门且准确率较高的人体姿态估计算法。

表 1 对比分析了其它 4 种算法和本文算法在 COCO 数据集上的检测结果,本文算法在 COCO 数据集上的平均准确率为 82.4%,召回率 89.3%,而 Personlab 算法和 PifPaf 算法平均准确率分别为 0.687 和 0.787,召回率分别为 0.754 和 0.864 均低于本文算法。此外,本文算法对 OKS=0.50 和 OKS=0.75 时,也分别计算了准确率与召回率,均高于其他 4 种。由于准确率和召回率越高,表明漏检和误检的关节点越少,证明本文提出的人体姿态估计算法能有效提高关节点检测与关联性能。 (下转第 146 页)