文章编号: 1671-6833(2021)03-0026-07

基于密集深度插值的 3D 人体姿态估计方法

陈梦婷, 王兴刚, 刘文予

(华中科技大学 电子信息与通信学院,湖北 武汉 430074)

摘 要: 3D 人体姿态估计是计算机视觉任务中一直非常具有挑战的任务。由于样本标注难度大,往往只能获得有限场景下的离散关键点数据,给三维的预测带来了更大的挑战。研究发现,虽然人体是一个非常灵活的结构,但是单个躯干可以看作刚体。这意味着当只知道躯干两端的深度时,整个躯干的深度都可以通过密集插值得到估计值。因此,提出了一种可以将每个躯干的密集深度插值特征图作为中间监督的方法。该特征图为深度的估计提供了更加密集、更加结构化的学习目标,而不仅仅是直接对离散关键点的深度进行回归。在数据集 Human3.6M 上的实验结果表明,该方法仅仅通过简单的网络结构,平均每个关节位置误差达到 50.9 mm。在数据集 MPI-INF-3DHP 上进行的跨域实验进一步证明了模型强大的泛化能力。

关键词: 3D 视觉; 人体姿态估计; 密集深度插值; 跨域泛化

中图分类号: TP391.41 **文献标志码:** A **doi:** 10.13705/j.issn.1671-6833.2021.03.005

0 引言

人体姿态估计一直是计算机视觉领域[1]中一个非常基础却又非常具有挑战性的任务。在给定图像或视频的情况下,预测人体关键点的 2D 或 3D 位置信息,这对于虚拟现实、增强现实、自动驾驶等需要空间推理的应用场景而言是至关重要的。得益于深度卷积神经网络(DCNN)的快速发展以及大规模手动注释的数据集的获取,目前在 2D 人体姿态估计方面已经取得了重大进展。

反观 3D 人体姿态估计的进展仍然有限,主要是由于在不受限制的环境中难以获得人体关节 3D 位置的真实标签。现有的数据集(例如 Human3.6M[2])是使用 Mocap 系统在受限的室内实验室环境中收集的,这样采集得到的数据集无论是在视角还是在光照和场景的变化上都比较单一。虽然深度卷积神经网络能够很好地拟合这类数据集,但将这样训练得到的模型运用到仅有 2D 标注的不受限的场景图片上时(例如 MPII[3]、MPI-INF-3DHP[4]),模型的表现往往不尽如人意。

研究发现,虽然人体是一个可以活动的结构,但是单个躯干(比如上臂、大腿等)可以近似看作是刚体结构。虽然数据集仅仅标注了关键点的 3D 信息,本文可以利用躯干两端的深度,通过密集插值估算出整个躯干的深度信息,从而构成密集深度插值特征图。本文将这个深度特征图作为模型训练的中间监督,这样可以为模型提供一个更加结构化的学习目标,而不仅仅是学习离散关键点的信息,从而有效提高模型的泛化能力,避免过拟合。而且在 3 个维度的学习过程中,深度学习往往是最具有难度的,通过密集深度特征图,可以让模型学习到结构化的深度信息,从而缓解因为遮挡、视觉变形带来的误差。

1 相关工作

1.1 2D 人体姿态估计

树形结构模型最早被用来解决 2D 人体姿态估计问题,比如 pictoral structures[5]和 mixtures of body parts[6],其主要思路是设计一个用于检测人体关节的一元项,加上用于模拟人体 2 个关节之间的成对关系的成对项。还有传统方法中建立四肢之间外观的对称性模型或是设计两臂之间的排斥边缘,以解决重复计数问题[7]。最近,DCNN 取

收稿日期: 2020-11-09; 修订日期: 2020-12-11

基金项目: 国家自然科学基金资助项目(61733007)

通信作者: 刘文予(1963—),男,湖南株洲人,华中科技大学教授,博士,博士生导师,主要从事人工智能、计算机视觉、多媒体通信与信息处理等研究,E-mail: liuwy@hust.edu.cn。

得了令人瞩目的进展[8]。相较于直接回归关键点的坐标[8]，目前更常见的做法是使用热力图，即以人体关节位置为中心的二维高斯生成的特征图作为模型回归的目标。常见的主干网络有 ResNet[9]、hourglass[10] 和 multi-stage 网络[11]。本文使用最新的 HRNet[12] 作为网络的主干架构。

1.2　3D 人体姿态估计

3D 人体姿态估计与 2D 人体姿态估计一直有很多相关之处。Lee 等[13] 首先研究了从相应的 2D 投影中来推断 3D 关键点的方法。后来的方法有的是利用最近临近算法来完善姿态推断[14]，有的是提取手工特征来完成回归[15]。

后来越来越多的研究致力于利用深度神经网络来完成这一任务。可以大致分为单阶段方法和两阶段方法。单阶段的方法希望可以直接由输入图像得到 3D 人体姿态的估计结果。Pavlakos 等[16] 提出了 3D 关节的体积表示，并使用了从粗粒度到精粒度的策略来迭代地精修预测结果。此类方法都需要具有相应 3D 标注的图像。由于缺乏带有 3D 标注的室外场景图像，这些方法往往会在跨域数据集上效果较差。Yang 等[17] 将 3D 姿态估计器看作是生成器，并使用对抗学习的方法生成令判别器无法区分的 3D 姿态，以保证预测结果结构上的真实性。而两阶段方法主要是先学习一个 2D 人体姿态估计的模型，再学习从 2D 到 3D 的映射模型。比如在 2D 人体姿态估计模型的后面加一个优化模型[18] 或者是回归模型[19-20] 来完成对 3D 姿态的估计。比如 Martinez 等[20] 引入了一种简单而有效的方法，可以仅通过对关键点的 2D 预测得到 3D 关键点的预测结果。Fang 等[21] 通过姿势语法网络进一步扩展了这种方法。这类方法往往能更好地泛化到其他室外场景数据集上。

2　密集插值姿态估计网络

2.1　密集深度插值

作为 3D 关键点任务检测，数据集只有离散的关键点的 3D 标注信息，所以很多方法仅仅通过 2D 的热力图作为中间特征，来帮助最后的 3D 回归。本文发现，虽然人体是非常灵活的结构，但是单独去看人体的某个躯干（比如左小臂、右大腿），可以近似地把它们看作一个刚体。因此，当仅仅只知道躯干两端点的深度信息时，可以近似估计出整个躯干的深度。

如图 1 所示，此处以一个小臂为例。P_w 和 P_e 代表关键点手腕 w（wrist）和手肘 e（elbow）的 2D 位置，它们构成第 m 个躯干。这两点的深度真实值分别为 $D_m(P_w)$ 和 $D_m(P_e)$。那么 P_w 和 P_e 连线上的任意点 P' 的深度 $D_m(P')$ 都可以通过线性插值进行估算：

$$\frac{D_m(P')-D(P_w)}{D_m(P_e)-D(P_w)}=\frac{\|P'-P_w\|_2}{\|P_e-P_w\|_2}。\tag{1}$$

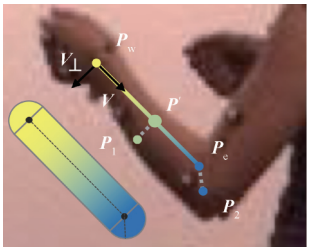


图 1　密集深度插值示意图

Figure 1　Diagram of dense depth interpretation map

不仅仅是两点连线上的点，本文对于位于躯干上的点 P_1 都可以给出估计深度，只要 P_1 满足：

$$\begin{cases}0\leqslant V(P_1-P_w)\leqslant\|P_e-P_w\|_2;\\ |V_\perp(P_1-P_w)|\leqslant r。\end{cases}\tag{2}$$

式中：r 代表躯干的半径；$V=(P_e-P_w)/\|P_e-P_w\|$。所有满足上述 2 个条件的点本文都认为是属于躯干的点，它的深度 $D_m(P_1)$ 等于 P_1 到 P_w 和 P_e 连线的垂足的点的深度 $D_m(P')$。

除了上述矩形空间，本文对关键点附近的区域点 P_2 也进行了深度估计：

$$\begin{cases}V(P_2-P_w)\geqslant\|P_e-P_w\|_2;\\ \|P_2-P_e\|_2\leqslant r。\end{cases}\tag{3}$$

$$\begin{cases}V(P_e-P_2)\geqslant\|P_e-P_w\|_2;\\ \|P_2-P_w\|_2\leqslant r。\end{cases}\tag{4}$$

所有满足式（3）范围内的点的深度等于 $D_m(P_e)$；所有满足式（4）范围内的点的深度等于 $D_m(P_w)$。最后得到的范围区域以及对应的预估深度图如图 1 所示。

每个躯干由一个单独的特征通道表示，本文采用一共有 16 个关键点组成的 15 个躯干，因此密集深度插值构成的目标特征共有 15 个通道，如图 2 所示。每个通道仅有部分属于躯干的点才有深度回归的目标，其他点因为没有目标值，所以在计算损失函数时不考虑。最后构造得到的目标特征图用 D 表示，它的第 m 个通道为 D_m，代表第 m 个躯干的连续深度分布。

2.2　辅助 2D 热力图

上述密集深度插值特征既包含了躯干在 2D 平面的位置信息，还包括了躯干的连续深度值。

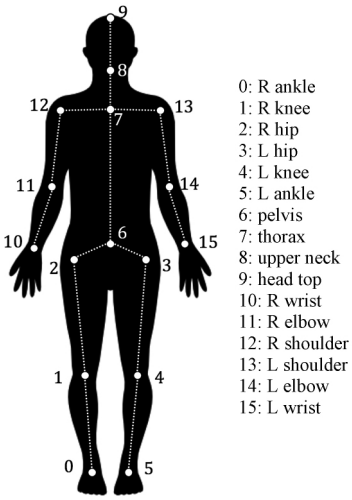
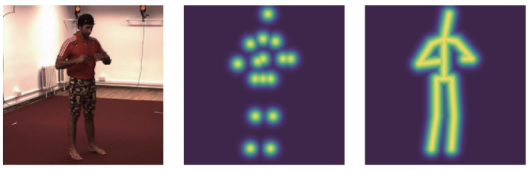


图 2 人体躯干示意图

Figure 2 Diagram of human body

但是相对而言学习起来比较困难。为了能够更好地学习拟合该特征图,本文用另外两个 2D 热力图作为辅助分支,如图 3 所示。



(a) 原图 (b) 关键点热力图 (c) 躯干热力图

图 3 辅助 2D 热力图示意图

Figure 3 Diagram of auxiliary 2D heat map

在关键点热力图中,每个关键点单独占一个通道。假设 $\boldsymbol{P}_k$ 是第 k 个点在图像中的真实位置,且 $\boldsymbol{P}_k \in \boldsymbol{R}^2$。那么第 k 个关键点在位置 $\boldsymbol{P}$ 的置信度为

$$S_k(\boldsymbol{P}) = \exp\left(-\frac{\|\boldsymbol{P}-\boldsymbol{P}_k\|_2}{\sigma^2}\right)。\tag{5}$$

其中,σ 控制山峰的陡峭程度。由此构造得到的辅助 2D 热力图如图 3(b)所示。

上述辅助 2D 热力图仅仅表征了关键点的 2D 位置,为了能更好地辅助躯干的深度图,本文构造了另一个代表躯干位置置信度的热力图。同样,本文以 $\boldsymbol{P}_w$ 和 $\boldsymbol{P}_e$ 代表关键点手腕 w(wrist)和手肘 e(elbow)的 2D 位置为例,它们构成第 m 个躯干。对于所有满足式(2)的点 $\boldsymbol{P}_1$ 属于第 m 个躯干的置信度为

$$Q_m(\boldsymbol{P}_1) = \exp\left(-\frac{|\boldsymbol{V}_\perp(\boldsymbol{P}_1-\boldsymbol{P}_w)|}{\sigma^2}\right)。\tag{6}$$

对于所有满足式(3)或式(4)的点 $\boldsymbol{P}_2$,它们的置信度分别为

$$\begin{cases} Q_m(\boldsymbol{P}_2) = \exp\left(-\dfrac{\|\boldsymbol{P}_2-\boldsymbol{P}_e\|_2}{\sigma^2}\right), & \text{当满足式(3)}; \\ Q_m(\boldsymbol{P}_2) = \exp\left(-\dfrac{\|\boldsymbol{P}_2-\boldsymbol{P}_w\|_2}{\sigma^2}\right), & \text{当满足式(4)}。 \end{cases}\tag{7}$$

由此构造得到的辅助热力图如图 3(c)所示。

2.3 整体网络结构

当获取了上述 3 个目标特征图后,网络的整体框架如图 4 所示。整个训练过程分为 2 个阶段。第一个阶段是输入图像到中间特征的训练。这里的 Backbone 使用的是 HRNet[12] 结构,本文的最后一个模块分成 3 个不同的分支,来分别预测 3 个特征图,之前的所有网络都是共享参数。对于关键点和躯干的热力图,本文使用的是均方误差(*MSE*)损失函数。辅助关键点热力图的损失函数为

$$L_s = \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(\hat{\boldsymbol{S}}_k^{(h,w)} - \boldsymbol{S}_k^{(h,w)}\right)^2。\tag{8}$$

式中:$\hat{\boldsymbol{S}}_k^{(h,w)}$ 代表预测的关键点热力图的第 k 个通道的坐标为(h,w)的点;H 和 W 分别为特征图的长和宽;K 为关键点的总个数。

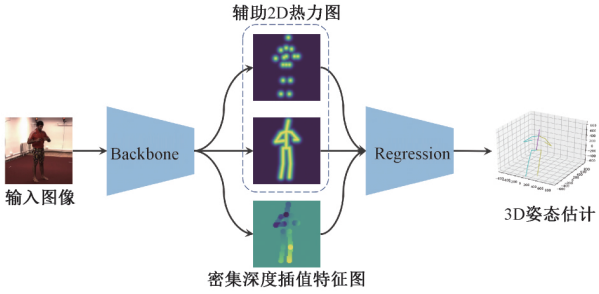


图 4 模型整体框架图

Figure 4 Diagram of model structure

因为对于躯干而言,不同的躯干的长度差异较大,为了避免因非 0 值的数量造成的差异,本文设置权重因子来平衡这种差异:

$$L_Q = \sum_{m=1}^{M} w_m \sum_{h=1}^{H}\sum_{w=1}^{W}\left(\hat{\boldsymbol{Q}}_m^{(h,w)} - \boldsymbol{Q}_m^{(h,w)}\right)^2;\tag{9}$$

$$w_m = \frac{1}{\sum_{h=1}^{H}\sum_{w=1}^{W}\boldsymbol{Q}_m^{(h,w)}}。\tag{10}$$

式中:w_m 为权重因子;$\hat{\boldsymbol{Q}}_m^{(h,w)}$ 代表预测的躯干热力图的第 m 个通道的坐标为(h,w)的点;M 为躯干总数。

对于密集深度插值特征图,因为只考虑躯干位置的深度,其他位置不参与损失函数的计算,所以通过躯干的辅助热力图对不考虑的点的损失函数设置为 0,并且也通过权重因子来平衡不同躯干的权重:

$$L_D = \sum_{m=1}^{M} w_m \sum_{h=1}^{H} \sum_{w=1}^{W} f\left(\widehat{\boldsymbol{D}}_m^{(h,w)}, \boldsymbol{D}_m^{(h,w)}\right) \boldsymbol{Q}_m^{(h,w)} 。 \tag{11}$$

式中: $\widehat{\boldsymbol{D}}_m^{(h,w)}$ 代表预测的深度特征图的第 m 个通道的坐标为 (h,w) 的点的值; $f(\cdot)$ 为 *smooth L*1 损失函数。第一阶段的训练损失函数为

$$L_1 = L_D + w_{2D}(L_S + L_Q) 。 \tag{12}$$

式中: w_{2D} 是辅助 2D 任务所占的权重。

第一阶段训练完成之后，用将第一阶段模型预测得到的 3 个输出作为输入，通过网络直接回归最后的 3D 姿态。使用的网络是由卷积层、最大池化层、ReLU 层以及全连接层组合得到。最后得到关键点的 3D 位置预测，采用两阶段的训练方式，主要是为了防止回归网络过拟合，中间监督失去作用，从而使网络的泛化性能变差。

3　实验结果

3.1　数据集

在 3 个最常见的人体姿态估计数据集上进行了实验。Human3. 6M [1] 数据集是最大的 3D 人体姿态估计数据集，它包含了 3.6×10^6 张图片，来自 11 个人。每人会表演 15 个日常动作，比如: 吃、坐下、行走和拍照等。数据集的 3D 姿态真实标签由 Mocap 系统获取，2D 姿态真实标签可以通过已知的摄像机内外部参数投影得到。参照 Human3. 6M 上的标准协议，评估指标为在对齐根关节深度后，所有关节的真实值与预测值的平均位置误差(*MPJPE*)，单位为 mm。

MPI-INF-3DHP [4] 数据集是最近提出的由 Mocap 系统构建的 3D 人体姿态数据集。本文仅使用该数据集的测试集，其中包含来自 6 个人的 7 个动作，共 2 929 张样本。本文用 *3DPCK*(阈值 150 mm) 和 *AUC* 两个指标来定量评估模型的泛化能力。

MPII [5] 数据集是 2D 人体姿态估计任务中使用最广泛的数据集之一。它包含从 YouTube 视频中收集的 2. 5 万张图像。数据集提供了 2D 标注，但没有 3D 的标注。因此，直接使用此数据集进行 3D 姿态估计训练是不可行的，故本文将此数据集用于多任务网络的训练。

3.2　实验结果

在目前最常用的 3D 人体姿态估计数据集 Human3. 6M 上进行了评估。和之前的许多方法一样，在第一阶段的训练过程中，联合 MPII 数据一起训练。因为 MPII 只有 2D 标注，所以只参与辅助 2D 分支的训练。详细的结果和对比如表 1 所示。可以看出，本文方法和之前的方法相比，结构更加清晰简单，而且具有更好的性能。

表 1　在 Human3. 6M 上的 *MPJPE* 比较结果

Table 1　Results of *MPJPE* on Human3. 6M　　mm

模型来源	出处	不同动作的 *MPJPE*							
		指导	讨论	吃	迎接	通话	拍照	造型	购买
文献 [22]	CVPR′17	89. 87	97. 57	89. 98	107. 87	107. 31	139. 17	93. 56	136. 09
文献 [19]	TPAMI′18	87. 36	109. 31	87. 05	103. 16	116. 18	143. 32	106. 88	99. 78
文献 [20]	CVPR′17	64. 98	73. 47	76. 82	86. 43	86. 28	110. 67	68. 93	74. 79
文献 [16]	CVPR′18	58. 55	64. 56	63. 66	62. 43	66. 93	70. 74	57. 72	62. 51
文献 [23]	ICCV′17	54. 82	60. 7	58. 22	71. 41	62. 03	65. 53	53. 83	55. 58
文献 [20]	ICCV′17	51. 80	56. 20	58. 10	59. 00	69. 50	78. 40	55. 20	58. 10
文献 [17]	CVPR′18	51. 50	58. 90	50. 40	57. 00	62. 10	**65. 40**	49. 80	52. 70
文献 [24]	ECCV′18	**43. 80**	51. 70	48. 80	53. 10	52. 20	74. 90	52. 70	**44. 60**
本文		45. 30	**47. 60**	**45. 50**	**48. 90**	**49. 70**	68. 90	**49. 60**	46. 40

模型来源	出处	不同动作的 *MPJPE*							
		坐着	坐下	吸烟	等待	遛狗	走路	走过	平均
文献 [22]	CVPR′17	133. 14	240. 12	106. 65	106. 21	87. 03	114. 05	90. 55	114. 18
文献 [19]	TPAMI′18	124. 52	199. 23	107. 42	118. 09	114. 23	79. 39	97. 70	79. 90
文献 [20]	CVPR′17	110. 19	172. 91	84. 95	85. 78	86. 26	71. 36	73. 14	88. 39
文献 [16]	CVPR′18	76. 84	103. 48	65. 73	61. 56	67. 55	56. 38	59. 47	66. 92
文献 [23]	ICCV′17	75. 20	111. 59	64. 15	66. 05	51. 43	63. 22	55. 33	64. 90
文献 [20]	ICCV′17	74. 00	94. 60	62. 30	59. 10	65. 10	49. 50	52. 40	62. 90
文献 [17]	CVPR′18	69. 20	85. 20	57. 40	58. 40	**43. 60**	60. 10	47. 70	58. 60
文献 [24]	ECCV′18	56. 90	**74. 30**	56. 70	66. 40	68. 40	47. 50	45. 60	55. 80
本文		**56. 30**	78. 90	**51. 20**	**48. 80**	46. 70	**40. 40**	**41. 30**	**50. 90**

3.3 跨域泛化结果

本文使用数据集 MPI-INF-3DHP 来验证模型到另一个全新的 3D 人体姿态估计数据集上的跨域迁移能力,该数据集的所有数据都不会参与训练过程,比较结果如表 2 所示。可以看出,通过密集插值特征图训练得到的模型具有更强的泛化迁移能力。

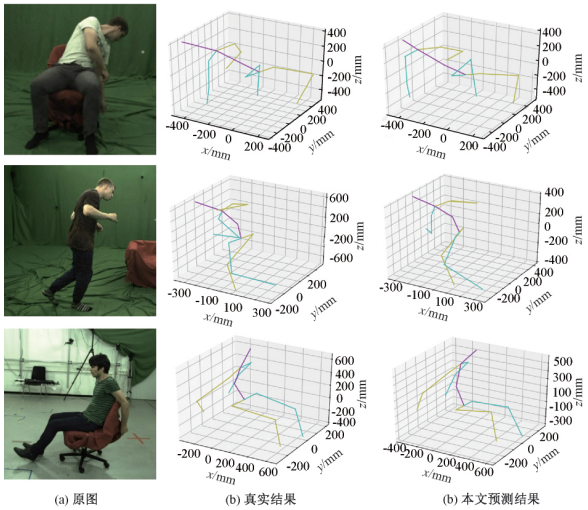
模型在数据集 MPI-INF-3DHP [4] 上的可视化结果如图 5 所示。可以看出,即使在出现物体遮挡或者姿态比较独特的时候,本文的模型也可以给出精确的结果。

表 2 在 MPI-INF-3DHP 上的跨域验证实验结果

Table 2 Results of domain transfer on MPI-INF-3DHP

mm

模型来源	出处	*3DPCK*	*AUC*
文献[4]	3DV'17	64.7	31.7
文献[23]	ICCV'17	69.2	32.5
文献[24]	CVPR'18	71.9	35.3
文献[18]	CVPR'18	69.0	32.0
文献[25]	ICCV'19	71.9	35.8
本文		**73.2**	**38.9**



(a) 原图 (b) 真实结果 (b) 本文预测结果

图 5 在数据集 MPI-INF-3DHP 上的可视化结果

Figure 5 Visualization on MPI-INF-3DHP

3.4 消融实验

首先比较了分两个阶段训练与单阶段联合训练的区别,实验结果如表 3 所示。可以看出,如果采用单一阶段的训练方式,在 Human3.6M 上的 *MPJPE* 结果会有细微提升,但是如用训练好的模型直接在数据集 MPI-INF-3DHP 做跨域验证时,*3DPCK* 和 *AUC* 都有大幅度下降,说明只有分两阶段训练,才能强制模型去学习有用的结构化信息,而不是直接去拟合离散关键点。这也进一步证明了本文所提出的密集深度插值特征图可以为模型带来更强的泛化能力。

表 3 不同训练方式在 Human3.6M 和 MPI-INF-3DHP 上的结果

Table 3 Results of different training strategy on Human3.6M and MPI-INF-3DHP mm

模型	*MPJPE*	*3DPCK*	*AUC*
单阶段	49.8	68.1	31.5
两阶段	50.9	73.2	38.9

4 结论

提出了一种基于线性插值的密集深度插值特征图作为 3D 人体姿态估计任务的中间监督,并通过两个辅助 2D 热力图来降低学习难度。通过在公认基准 Human3.6M 上的实验证明了该特征图的有效性和简洁性。并通过在 MPI-INF-DHP 上的跨域验证实验展示了模型强大的泛化迁移能力。由此可以看出,用结构化的深度信息作为学习目标可以有效地提高模型的性能。这种结构化也可以直接拓展到整个 3D 空间,将这种插值结构信息的作用发挥到最大,这也是本文未来的研究目标之一。

参考文献:

[1] 杨忠明,李子龙,胡音文,等. 一种前景提取的行人模式识别检测算法[J]. 郑州大学学报(工学版),2019,40(5):91-96.

[2] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6M: large scale datasets and predictive methods for 3D human sensing in natural environments [J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 36(7): 1325-1339.

[3] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: new benchmark and state of the art analysis [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2014: 3686-3693.

[4] MEHTA D, RHODIN H, CASAS D, et al. Monocular 3D human pose estimation in the wild using improved CNN supervision [C]// 7th IEEE International Conference on 3D Vision, 3DV. Piscataway: IEEE, 2017: 506-516.

[5] PISHCHULIN L, ANDRILUKA M, GEHLER P, et al. Poselet conditioned pictorial structures [C]// 26th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 588-595.

[6] YANG Y, RAMANAN D. Articulated pose estimation with flexible mixtures-of-parts [C]// Proceedings of the IEEE Computer Society Conference on Computer

Vision and Pattern Recognition. Piscataway: IEEE, 2011: 1385-1392.

[7] FERRARI V, MARÍN-JIMÉNEZ M, ZISSERMAN A. 2D human pose estimation in TV shows [J]. Statistical and geometrical approaches to visual motion analysis, 2009, 5064: 128-147.

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

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

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

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

[12] CARREIRA J, AGRAWAL P, FRAGKIADAKI K, et al. Human pose estimation with iterative error feedback [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 4733-4742.

[13] LEE H J, CHEN Z. Determination of 3D human body postures from a single view [J]. Computer vision, graphics, and image processing, 1985, 30(2): 148-168.

[14] GUPTA A, MARTINEZ J, LITTLE J J, et al. 3D pose from motion for cross-view action recognition via non-linear circulant temporal encoding [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2014: 2601-2608.

[15] ROGEZ G, RIHAN J, RAMALINGAM S, et al. Randomized trees for human pose detection [C] //2008 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2008: 1-8.

[16] PAVLAKOS G, ZHOU X W, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 1263-1272.

[17] YANG W, OUYANG W L, WANG X L, et al. 3D human pose estimation in the wild by adversarial learning [C] // Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5255-5264.

[18] ZHOU X W, ZHU M, PAVLAKOS G, et al. MonoCap: monocular human motion capture using a CNN coupled with a geometric prior [J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 41(4): 901-914.

[19] TOME D, RUSSELL C, AGAPITO L. Lifting from the deep: convolutional 3D pose estimation from a single image [C] //2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5689-5698.

[20] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3D human pose estimation [C] //Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 2640-2649.

[21] FANG H S, XU Y L, WANG W G, et al. Learning pose grammar to encode human body configuration for 3D pose estimation [EB/OL]. (2017-10-17) [2020-10-30]. https://arxiv.org/abs/1710.06513.

[22] CHEN C H, RAMANAN D. 3D human pose estimation = 2D pose estimation + matching [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5759-5767.

[23] ZHOU X, HUANG Q, SUN X, et al. Towards 3D human pose estimation in the wild: a weakly-supervised approach [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 398-407.

[24] WANG J, HUANG S L, WANG X C, et al. Not all parts are created equal: 3D pose estimation by modeling bi-directional dependencies of body parts [C] //Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2019: 7771-7780.

[25] LEE K, LEE I, LEE S. Propagating LSTM: 3D pose estimation based on joint interdependency [C] // European Conference on Computer Vision-ECCV 2018. Berlin: Springer, 2018: 123-141.

Dense Depth Interpolation for 3D Human Pose Estimation

CHEN Mengting, WANG Xinggang, LIU Wenyu

(School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: The 3D human pose estimation is a challenging task in computer vision. Due to the difficulty of annotation, only some disperse key-point data form limited scenes are available, which makes 3D prediction a big challenge. In this paper, the human body is deemed as a flexible structure, but a specific limb can be viewed as a rigid-body. Given depths of two points on both ends, the depths of the whole limb can be estimated by dense interpretation. Therefore, this paper proposes a method that can take the dense depth interpretation feature map as middle supervision. It provides a denser and more structured target, instead of regression for disperse key-points directly. The *MPJPG* on Human3.6M reaches 50.9 mm with only a simple network structure. The cross-domain experiments on dataset MPI-INF-3DHP further show the generalization ability of the proposed method.

Key words: 3D vision; human pose estimation; dense depth interpolation; cross-domain generalization

(上接第25页)

Multi-controller Deployment Strategy Based on Delay and Load Balancing

LIU Zhenpeng[1,2], WANG Xinpeng[1], LI Ming[1], REN Shaosong[1], LI Xiaofei[2]

(1. School of Electronic Information Engineering, Hebei University, Baoding 071002, China; 2. Center for Information Technology, Hebei University, Baoding 071002, China)

Abstract: To address the time delay and load balancing problems faced by multiple controllers deployed in the software definition network (SDN), in this paper, a multi-controller placement algorithm is proposed to reduce the time delay between controllers, and improve the network performance on the basis of load balancing. Aiming at the slow convergence speed of traditional particle swarm optimization algorithm, this paper proposes an improved particle swarm optimization algorithm to deploy the SDN controller. The improved particle swarm optimization algorithm is used to deploy the SDN controller to minimize the propagation delay between the switch and the controller while considering the load balance of the controller. The simulation results show that the improved particle swarm optimization algorithm for controller deployment can guarantee high load balancing performance and the better overall network performance by acquire fitness about 0.05. And compared with the traditional particle swarm optimization algorithm, the improved particle swarm optimization algorithm can improve the convergence speed of the whole network about 6.3% with lower time delay.

Key words: software defined network; controller placement; latency; load balancing; particle swarm optimization