多视图快速鲁棒的多人 3D 姿态估计

董峻廷 21821102

Abstract

本文解决的问题是从几个标定好的相机中估计多人 *3D* 姿态。该问题的主要挑战是找到杂乱且不完整的 *2D* 姿态之间的交叉视图对应关系。大多数先前的方法通过使用图形结构模型 *(pictorial structure model)* 在 *3D* 中直接推理来解决该挑战，由于巨大的状态空间，该模型效率低。我们提出了一种快速而鲁棒的方法来解决这个问题。我们的主要思想是使用多路径匹配算法在所有视图中聚类检测到的 *2D* 姿势。每个得到的聚类在不同视图上编码同一人的 *2D* 姿势以及关键点上的一致对应，从中可以有效地推断出每个人的 *3D* 姿势。所提出的基于凸优化的多路径匹配算法对于丢失和错误检测是有效且鲁棒的，即使不知道场景中的人数。此外，我们结合几何和外观线索进行交叉视图匹配。所提议的方法比最先进的算法有显著性能提升（分别为 96.3% *vs.* 90.6% 和 96.9% *vs.* 88%，分别在 *Campus* 和 *Shelf* 数据集上*)*, 而且速度快适合实时应用。

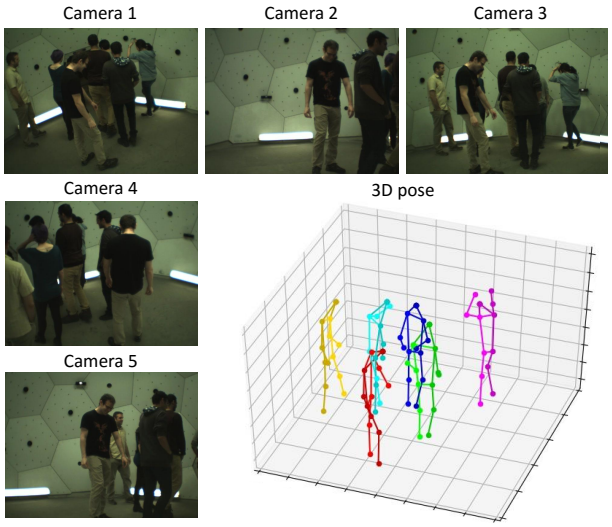

图 1: 本文提出了一种新方法，可以从几个相机视图快速，稳健地恢复多人的 3D 姿态。主要挑战是在多个视图之间建立一致的 2D 观察对应，例如图像中的 2D 人体关键点，很可能含有噪声且不完整。

1. 引言

从视频中恢复 3D 人体姿态一直是计算机视觉中长期存在的问题，其具有诸如人机交互，视频监视等各种应用。本文重点关注场景中有多个人的情况，通过几个标定好的相机作为输入（图1）。虽然人体的多视图重建方面已经取得了显着的进步，但是在一些具有挑战性的场景，比如多个人在拥挤的场景中彼此交互，存在大量遮挡, 还是只有比较少的工作。

现有方法通常使用两个阶段的方法来解决该问题。第一阶段在每个的 2D 视图中检测人体关键点，在第二阶段中聚类来重建 3D 姿态。鉴于基于深度学习的 2D 关键点检测技术已经取得了显着的性能 [6, 17]，余下的挑战是找到检测到的关键点之间的交叉视图对应关系以及它们属于哪个人。大多数先前的方法 [1, 2, 13, 9]采用 3D 图形结构（3DPS）模型，通过推理 3D 中与 2D 检测在几何上兼容的所有候选点，隐式地解决对应问题。然而，由于巨大的状态空间，这种基于 3DPS 的方法在计算上是昂贵的。此外，特别是当相机数量较少时，它不够稳定，因为它仅使用多视图几何结构来跨视图匹配 2D 检测，或者换句话说，忽略了人的外观线索。

在本文中，我们提出了一种用于多人 3D 姿态估计的新方法。所提出的方法通过在多个视图之间匹配检测到的 2D 姿态来解决人的对应问题，产生 2D 姿态的类，其中每个类包括不同视图中的同一人的 2D 姿态。然后，可以针对每个人与匹配的 2D 姿态分开地推断 3D 姿态，这比由于减小的状态空间的多个姿态的联合

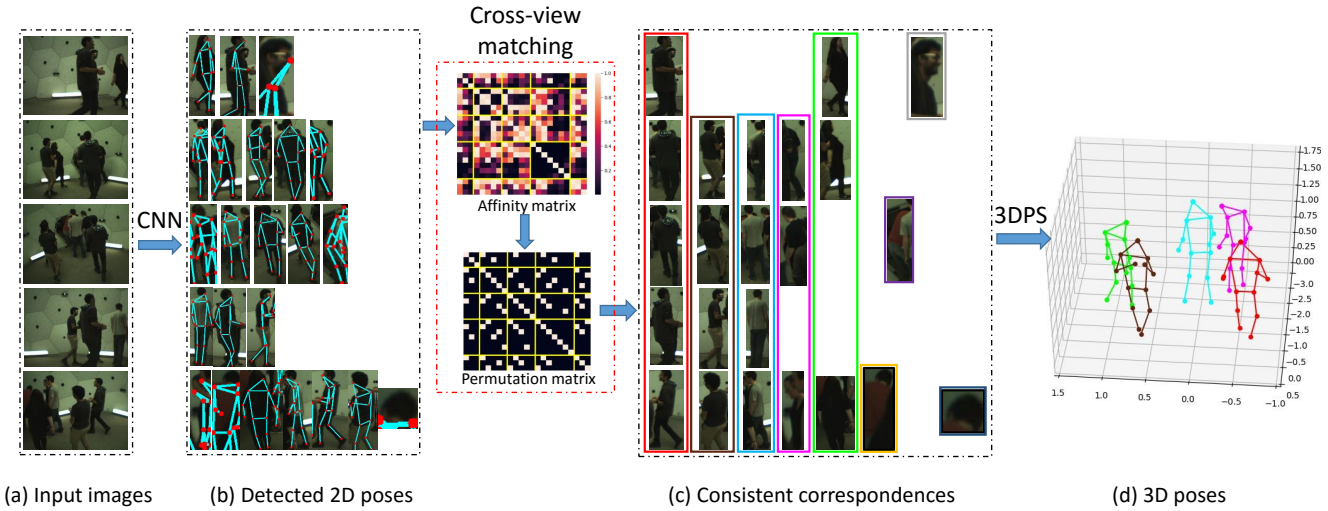

图 2: 方法概述。给定来自几个标定好的相机的图像 (a)，使用现成的人体姿态检测器在每个视图中产生 2D 边界框和相应的 2D 姿态，这可能是不准确和不完整的（b）。然后，通过本文提出的多视图匹配算法对检测到的边界框进行聚类。每个结果簇包括不同视图中同一个人的边界框 (c)。在其他视图中没有匹配的单一边界框被视为错误检测并被丢弃。最后，从相应的边界框和相关的 2D 姿态中重建每个人的 3D 姿态 (d)。

推断快得多。

但是，在多个视图中匹配 2D 姿态是非常具有挑战性的。一种典型的方法是使用极几何约束来验证两个 2D 姿态是否是每对视图的相同 3D 姿态的投影 [15]。但是这种方法可能由于以下原因而失败。首先，检测到的 2D 姿态通常由于严重的遮挡和截断而不准确，如图 2(b) 所示，这使得几何验证变得困难。其次，分别匹配每对视图可能产生违反回路一致性约束的不一致对应，即两个视图中的两个对应姿态可以与另一视图中的不同人匹配。这种不一致导致不正确的多视图重建。最后，如图 2所示，不同的人出现在不同的视图中，总人数未知，这给匹配问题带来了额外的困难。

我们提出了一种多路径匹配算法来解决上述挑战。我们的主要思想是：（i）将 2D 姿态之间的几何一致性与其相关图像块之间的外观相似性进行结合，以减少匹配不确定性，以及（ii）同时利用回路一致性约束解决所有视图的匹配问题以利用多个相机的信息并产生全局一致的对应。将匹配问题表示为凸优化问题，并提出一种有效的算法来解决该优化问题。

2. 相关工作

无标记运动捕捉已经在计算机视觉中进行了十年的研究。关于这个问题的早期工作旨在通过多视图序列跟踪人体的三维骨架或几何模型 [20, 21, 8]。这些基于跟踪的方法需要在第一帧中进行初始化，并且易于出现局部最优和跟踪失败。因此，最近的工作通常基于自下而上的方案，其中 3D 姿态是从图像中检测到的 2D 特征重建的：[19, 5, 18]。最近的工作 [14]通过将统计体模型与基于深度学习的 2D 姿态检测器相结合，显示了显着的结果。

在本文中，我们专注于多人 3D 姿态估计。以前的大多数工作都是基于 3DPS 模型，其中节点代表身体关节的 3D 位置，边编码它们之间的成对关系 [1, 12, 2, 13, 9]。每个关节的状态空间通常是表示离散化 3D 空间的 3D 网格。关节位于某个位置的可能性由应用于所有 2D 视图的关节检测器给出，关节之间的成对关系由骨架约束给出 [1, 2]或在 2D 视图中检测到的身体部位 [13, 9]。然后，通过最大后验估计共同推断出多人的 3D 姿态。

由于所有人的所有身体关节被同时考虑，整个状态空间巨大，导致推理计算量很大。这种方法的另一个限制是它只使用多视图几何来连接 2D 观察，这对相机的设置很敏感。因此，当视图数量减少时，此方法的性

能会显着下降 [13]。最近的工作 [15]建议在视图之间匹配 2D 姿态，然后从属于同一个人的 2D 姿态重建 3D 姿态。但它仅利用极几何来匹配每对视图的 2D 姿态，并忽略多个视图之间的回路一致性约束，这可能导致不一致的对应关系。

3. 方法

Figure 2概述了我们的方法。首先，采用现成的 2D 人体姿态检测器来生成每个视图中人物的边界框和 2D 关键点位置（Section 3.1）。给定有噪声的 2D 检测，提出了一种多路径匹配算法来建立跨视图检测到的边界框的对应关系并消除错误检测（Section 3.2）。最后，3DPS 模型用于从相应的 2D 边界框和关键点（Section 3.3）重建每个人的 3D 姿态。

3.1. 2D 人体姿态检测

我们采用最近提出的级联金字塔网络 [7]在 MSCOCO [16]数据集上进行训练，以便在图像中进行 2D 姿态检测。级联金字塔网络由两个阶段组成：GlobalNet 大致估计人类姿态，而 RefineNet 则提供最佳人体姿态。尽管它在基准测试中具有最先进的性能，但检测的结果依然可能非常杂乱，如图2(b) 所示。

3.2. 多视图对应

在重建 3D 姿态之前，检测到的 2D 姿态应该跨视图匹配. 为了解决这个问题，我们需要 1）一个适当的度量来衡量两个 2D 边界框属于同一个人的可能性，和 2）一个匹配算法来建立跨多个视图的边界框的对应关系。特别是，匹配算法不应该对场景中的真实人数进行任何假设。此外，匹配算法的输出应该是回路一致的，即两个图像中的任何两个对应的边界框应该对应于另一个图像中的相同的边界框。

问题陈述： 在详细介绍我们的方法之前，我们先简要介绍一下符号的含义。假设场景中有 V 相机，在视角 i 中检测到 p_i 个边界框。对于一对视图 (i,j)，可以在视图 i 和视图 j 中的两组边界框之间计算相似性分数。我们使用 $\boldsymbol{A}_{ij} \in \mathbb{R}^{p_i \times p_j}$ 来表示相似性矩阵，其元素代表相似性值。在两组边界框之间的对应关系由部分置换矩阵 $\boldsymbol{P}_{ij} \in \{0,1\}^{p_i \times p_j}$ 表示，且满足双随机约束：

$$\mathbf{0} \leq \boldsymbol{P}_{ij}\mathbf{1} \leq \mathbf{1}, \mathbf{0} \leq \boldsymbol{P}_{ij}^T\mathbf{1} \leq \mathbf{1}. \tag{1}$$

The problem is to take $\{\boldsymbol{A}_{ij}|\forall i,j\}$ as input and output the optimal $\{\boldsymbol{P}_{ij}|\forall i,j\}$ 最大化相应的相似性值，并且在多个视图中也是回路一致的。

相似性矩阵： 我们结合外观相似性和极几何约束来计算边界框之间的相似性值。

首先，我们采用预先训练的行人重视别网络来提取边界框的外观特征。具体来说，我们通过 [22]中提出的行人重视别模型处理每个边界框的裁剪图像，并从“pool5”图层中提取特征向量作为每个边界框的特征描述。然后，我们计算边界框对的特征描述之间的欧几里德距离，并使用 sigmoid 函数将距离映射到 $(0,1)$ 中的值作为此边界框对的外观相似性值。

除了外观之外，关联两个边界框的另一个重要线索是它们相关的 2D 姿态应该是几何一致的。具体地，相应的 2D 关节位置应满足极几何约束。假设 $\boldsymbol{x} \in \mathbb{R}^{N\times 2}$ 表示由 N 关节组成的 2D 姿态。然后，两个视图中 $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ 之间的几何一致性可以通过以下距离来度量：

$$D_g(\boldsymbol{x}_i,\boldsymbol{x}_j) = \frac{1}{2N}\sum_{n=1}^{N} d_g(\boldsymbol{x}_i^n, \boldsymbol{L}_{ij}(\boldsymbol{x}_j^n)) + d_g(\boldsymbol{x}_j^n, \boldsymbol{L}_{ji}(\boldsymbol{x}_i^n)),$$

其中 $\boldsymbol{x}_i^n$ 表示姿态 i，$\boldsymbol{L}_{ij}(\boldsymbol{x}_j^n)$ 的 n-th 关节的 2D 位置从另一个视图与 $\boldsymbol{x}_j^n$ 相关联的极线，$d_g(\cdot,\boldsymbol{l})$ l 的点到线距离。距离 D_g 也使用 sigmoid 函数作为最终几何相似性度分数映射到 $(0,1)$ 中的值。

基于一对正确检测和匹配的 2D 姿态必须满足几何约束（D_g 很小）的事实，我们将两个相似性度矩阵组合如下：

$$\boldsymbol{A}_{ij}(\cdot) = \begin{cases} \sqrt{\boldsymbol{A}_{ij}^a(\cdot) \times \boldsymbol{A}_{ij}^g(\cdot)}, & \text{if } D_g \leq th, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

其中 $\boldsymbol{A}_{ij}(\cdot)$，$\boldsymbol{A}_{ij}^a(\cdot)$ 和 $\boldsymbol{A}_{ij}^g(\cdot) \in [0,1]$ 分别表示视图对 (i,j) 的相似性矩阵，外观相似性矩阵和几何相似性矩阵的值。th 表示阈值。

多路径匹配与回路一致性： 如果只有两个视图匹配，可以简单地最大化 $\langle \boldsymbol{P}_{ij}, \boldsymbol{A}_{ij} \rangle$ 并使用匈牙利算法找到最佳匹配。但是当存在多个视图时，如果还是每对视图单

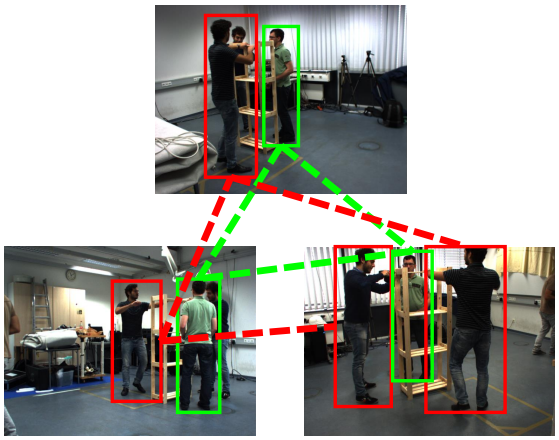

图 3: 回路一致性的例子。绿线表示一组一致的对应关系，红线表示一组不一致的对应关系。

独解决匹配问题会忽略回路一致性约束，并可能导致不一致的结果。图 3显示了一个例子，其中红色的对应关系不一致，绿色的对应关系是回路一致的，因为它们形成一个闭合回路。

我们利用 [11]中的结果来解决这个问题。假设在所有视图中所有检测到的边界框 $m = \sum_{i=1}^{V} p_i$ 之间的对应关系由 $\boldsymbol{P} \in \{0,1\}^{m\times m}$:

如果只有两个视图匹配，可以简单地最大化 $\langle \boldsymbol{P}_{ij}, \boldsymbol{A}_{ij} \rangle$ 并找到匈牙利算法的最佳匹配。但是当存在多个视图时，为每对视图单独解决匹配问题会忽略回路一致性约束，并可能导致不一致的结果。图 3显示了一个例子，其中红色的对应关系不一致，绿色的对应关系是回路一致的，因为它们形成一个闭合回路。

我们利用 [11]中的结果来解决这个问题。假设在所有视图中所有 $m = \sum_{i=1}^{V} p_i$ 检测到的边界框之间的对应关系 $\boldsymbol{P} \in \{0,1\}^{m\times m}$:

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P}_{11} & \boldsymbol{P}_{12} & \cdots & \boldsymbol{P}_{1n} \\ \boldsymbol{P}_{21} & \boldsymbol{P}_{22} & \cdots & \boldsymbol{P}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{P}_{n1} & \cdots & \cdots & \boldsymbol{P}_{nn} \end{pmatrix}, \tag{3}$$

其中 $\boldsymbol{P}_{ii}$ 应该是单位阵。然后，可以证明当且仅当

$$\text{rank}(\boldsymbol{P}) \leq s,\ \boldsymbol{P} \succeq 0, \tag{4}$$

满足回路一致性约束, 其中 s 是场景中的真实人数。由于 s 事先是未知的，我们通过最小化以下目标函数来估计低秩和正半定矩阵 $\boldsymbol{P}$:

$$\begin{aligned} f(\boldsymbol{P}) &= -\sum_{i=1}^{n}\sum_{j=1}^{n}\langle \boldsymbol{A}_{ij}, \boldsymbol{P}_{ij}\rangle + \lambda \cdot \text{rank}(\boldsymbol{P}), \\ &= -\langle \boldsymbol{A}, \boldsymbol{P}\rangle + \lambda \cdot \text{rank}(\boldsymbol{P}), \end{aligned} \tag{5}$$

其中 $\boldsymbol{A}$ 是所有 $\boldsymbol{A}_{ij}$ 的级联，类似于(3)中的形式，λ 表示低秩约束的权重。

优化: 为了使优化易于处理，我们做了适当的松弛。我们最小化核范数 $\|\boldsymbol{P}\|_*$，而不是最小化秩，这是因为它是秩最紧密的凸近似 [10]。我们把 P 当做一个值为 $[0,1]$ 的实数矩阵来替换 $\boldsymbol{P}$ 上的整数约束:

$$0 \leq \boldsymbol{P} \leq 1, \tag{6}$$

这是匹配算法的常见做法。我们去除了半正定约束，只要求 $\boldsymbol{P}$ 对称:

$$\boldsymbol{P}_{ij} = \boldsymbol{P}_{ji}^{T},\ \ 1 \leq i,j \leq n, i \neq j, \tag{7}$$

$$\boldsymbol{P}_{ii} = \boldsymbol{I}_{p_i},\quad 1 \leq i \leq n. \tag{8}$$

最后，我们解决了以下优化问题:

$$\begin{aligned} \min_{\boldsymbol{P}}\ & -\langle \boldsymbol{A}, \boldsymbol{P}\rangle + \lambda\|\boldsymbol{P}\|_*, \\ \text{s.t.}\quad & \boldsymbol{P} \in \mathcal{C}, \end{aligned} \tag{9}$$

其中 $\mathcal{C}$ 表示满足约束的可行域集合(1)，(6)，(7)和(8)。注意到(9)中的问题是凸的，我们使用 ADMM[4]来解决它。

3.3. 3D 姿态重建

在得到了不同视图中估计的同一人的 2D 姿态，我们重建每个人的 3D 姿态。这可以通过三角测量简单地完成，但 2D 姿态估计中的误差可能在很大程度上降低重建精度。为了在 2D 姿态估计中完全整合不确定性并将结构先验结合到人体骨架上，我们利用 3DPS 模型来估计 3D 姿态, 这是一个比较成熟的算法, 这里不展开讲.

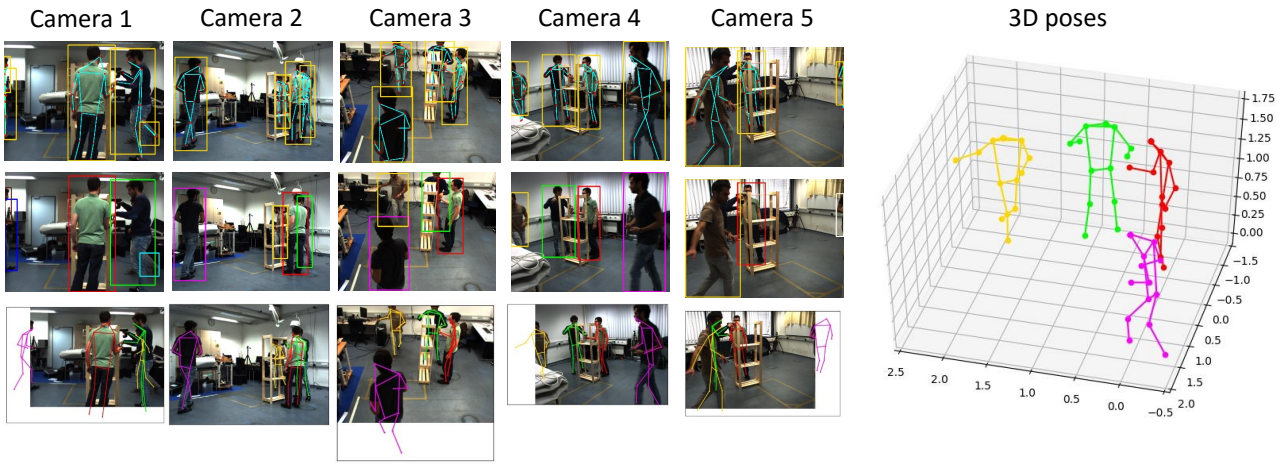

图 4: Shelf（顶部）数据集的定性结果。第一行显示 2D 边界框和姿态检测。第二行显示了我们的匹配算法的结果，其中颜色表示跨视图的边界框的对应关系。第三行显示估计的 3D 姿态的 2D 投影。

4. 实验

我们在两个公共数据集（包括室内和室外场景）上评估所提出的方法，并将其与之前的工作进行比较。

4.1. 数据集

Campus[1]: 它是由三个人在室外环境中相互交互组成的数据集，由三个标定好的摄像头捕获。我们遵循与以前的工作相同的评估方式 [1, 3, 2, 9]并使用正确估计部分（PCP）的百分比来测量身体部位的 3D 位置的准确性。

Shelf[1]: 与 Campus 相比，这个数据集更加复杂，包括四个人在近距离拆卸货架。它们周围有五个标定好的相机，但每个视图都有严重的遮挡。评估方法遵循先前的方法，评估指标也是 3D PCP。

4.2. 与最先进的技术比较

Campus 和 Shelf 数据集的结果显示在表1中。请注意，在我们的方法中使用的 2D 姿态检测器 [7]和 reID network [22]是已发布的预先训练好的模型，而不在评估的数据集上进行任何调整。即使使用通用模型，我们的方法也大大优于最先进的方法。特别是，我们的方法显着提高了 Campus 数据集中 Actor 3 和 Shelf 数据集中 Actor 2 的性能，该数据集遭受严重遮挡。

Campus	Actor 1	Actor 2	Actor 3	Average
Belagiannis *et al.* [1]	82.0	72.4	73.7	75.8
Belagiannis *et al.* [3]	83.0	73.0	78.0	78.0
Belagiannis *et al.* [2]	93.5	75.7	84.4	84.5
Ershadi-Nasab *et al.* [9]	94.2	92.9	84.6	90.6
我们	**97.6**	**93.3**	**98.0**	**96.3**
Shelf	Actor 1	Actor 2	Actor 3	Average
Belagiannis *et al.* [1]	66.1	65.0	83.2	71.4
Belagiannis *et al.* [3]	75.0	67.0	86.0	76.0
Belagiannis *et al.* [2]	75.3	69.7	87.6	77.5
Ershadi-Nasab *et al.* [9]	93.3	75.9	94.8	88.0
我们	**98.8**	**94.1**	**97.8**	**96.9**

表 1: Campus 和 Shelf 数据集的定量比较。数字是估计部位的（PCP）百分比。其他方法的结果取自各自的论文。

4.3. 定性评估

Figure 4显示了在 Shelf 数据集上提出的方法的一些代表性结果。将不准确的 2D 检测作为输入，我们的方法能够建立跨视图的对应关系，自动识别场景中的人数，最后重建他们的 3D 姿态。通过将 3D 姿态投影回 2D 视图而获得的最终 2D 姿态估计也比原始检测更精确。

5. 总结

在本文中，我们提出了一种新的多视图 3D 姿态估计方法，可以快速鲁棒地恢复一群人的 3D 姿态，只需几台标定好的相机。与之前基于 3DPS 的方法相比，我们的关键想法是使用多路径匹配算法来聚类检测到的 2D 姿态，以减少 3DPS 模型的状态空间，从而提高效率和鲁棒性. 实验结果验证了所提出的多路径匹配算法的有效性，该算法利用几何和外观线索的组合以及用于跨多个视图匹配 2D 姿态的回路一致性约束。

参考文献

[1] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures for multiple human pose estimation. In *CVPR*, 2014. 1, 2, 5

[2] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *T-PAMI*, 38(10):1929–1942, 2016. 1, 2, 5

[3] V. Belagiannis, X. Wang, B. Schiele, P. Fua, S. Ilic, and N. Navab. Multiple human pose estimation with temporally consistent 3d pictorial structures. In *ECCV workshop*, 2014. 5

[4] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011. 4

[5] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, 2013. 2

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1

[7] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. *CVPR*, 2018. 3, 5

[8] A. Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient convnet-based marker-less motion capture in general scenes with a low number of cameras. In *CVPR*, 2015. 2

[9] S. Ershadi-Nasab, E. Noury, S. Kasaei, and E. Sanaei. Multiple human 3d pose estimation from multiview images. *Multimedia Tools and Applications*, 77(12):15573–15601, 2018. 1, 2, 5

[10] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002. 4

[11] Q.-X. Huang and L. Guibas. Consistent shape maps via semidefinite programming. In *Proceedings of the Eleventh Eurographics/ACMSIGGRAPH Symposium on Geometry Processing*, pages 177–186. Eurographics Association, 2013. 4

[12] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 2

[13] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. S. Godisart, B. Nabbe, I. Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *T-PAMI*, 2017. 1, 2, 3

[14] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 2

[15] A. Kadkhodamohammadi and N. Padoy. A generalizable approach for multi-view 3d human pose regression. *CoRR*, abs/1804.10462, 2018. 2, 3

[16] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[17] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1

[18] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 2017. 2

[19] L. Sigal, M. Isard, H. W. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *IJCV*, 98(1):15–48, 2012. 2

[20] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *CVPR*, 2010. 2

[21] A. Yao, J. Gall, L. V. Gool, and R. Urtasun. Learning probabilistic non-linear latent variable models for tracking complex activities. In *NIPS*, 2011. 2

[22] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang. Camera style adaptation for person re-identification. In *CVPR*, 2018. 3, 5