# Multi-Person Hierarchical 3D Pose Estimation in Natural Videos

Renshu Gu, Gaoang Wang, Zhongyu Jiang, and Jenq-Neng Hwang, *Fellow, IEEE*

***Abstract*— Despite the increasing need of analyzing human poses on the street and in the wild, multi-person 3D pose estimation using monocular static or moving camera in real-world scenarios remains a challenge, either requiring large-scale training data or high computation complexity due to the high degrees of freedom in 3D human poses. We propose a novel scheme to effectively track and hierarchically estimate 3D human poses in natural videos in an efficient fashion. Without the need of using labelled 3D training data, we formulate torso estimation as a Perspective-N-Point (PNP) problem, and limb pose estimation as an optimization problem, and hierarchically structure the high dimensional poses to efficiently address the challenge. Experiments show good performance and high efficiency of multi-person 3D pose estimation on real-world videos, including street scenarios and various human daily activities from fixed and moving cameras, resulting in great new opportunities to understand and predict human behaviors.**

***Index Terms*— 3D human pose estimation, monocular camera, hierarchical, human tracking, visual odometry, perspective-n-point (PNP).**

## I. Introduction

ANALYZING human behaviors [46], [47] is one of the most popular research topics in recent years. 3D human pose estimation is central to analyzing human behaviors. Existing 3D human pose estimation calls for large-scale training data or high computation complexity, due to the high degrees of freedom in 3D human poses. In recent years, there have been many reports on 3D human pose estimation in the experimental setting. However, analysis is still lacking in natural videos for both fixed and moving cameras. To address the challenge in real-world applications, this paper proposes a novel and efficient scheme to recover and track 3D full-body human poses for multiple people from a monocular fixed or moving camera.

A popular solution of estimating 3D human pose estimation are deep learning methods that use powerful training from 3D motion capture (MoCap) data [18], [19], [28], [30]. Yet, these 3D pose estimation methods cannot be easily applicable for videos in the wild, since 3D MoCap training data are typically acquired in the indoor controlled environments. The lack of large-scale training data in the wild for fixed or moving cameras becomes a bottleneck. Moreover, existing methods might over-fit to sparse camera settings and bear poor generalization capabilities. Even with adequate training data, it is unclear how the space of 3D poses can be uniformly sampled. In our research, we thus decide not to rely on 3D MoCap ground truth training data.

Moreover, many training methods focus on single-person 3D pose estimation with the subject being at the center of the image, which makes it easy to associate estimated single-person body joints along the time for subsequent action and behavior analyses. On the other hand, it is much harder to handle multi-person 3D pose estimation where people can be interacting and occluding one another in real-world images or videos. Also, temporal information is still not well exploited in many of the existing works. These drawbacks can lead to a performance drop for natural video human pose estimation.

In this paper, we propose an efficient method to address the challenges encountered in multi-person 3D human pose estimation in natural videos from fixed or moving cameras. Our proposed method allows to efficiently reconstruct 3D human poses for multiple people in monocular image sequences with arbitrary camera motion. Unlike existing methods that feature high degrees of freedom (DoFs) in pose space, we structure the pose space in a hierarchical fashion to tackle the problem efficiently. It is achieved by utilizing recent advances in 2D pose estimation, i.e. OpenPose [1], associating and tracking multiple people to use temporal information, and then estimating each person's 3D human poses hierarchically with body geometric constraints. When estimating each person's poses, we apply a prior flexible human model that contains bone lengths of all human body parts, which can be optimized. Instead of trying to solve all the poses in high dimensions simultaneously, we first estimate the torso pose, and then estimate limb poses in a hierarchical fashion. Using 2D joints of multiple people from OpenPose [1] as an intermediate step, our method does not need to crop bounding boxes, and is robust to position changes. We demonstrate that our method can qualitatively produce smooth and natural 3D human poses on real-world datasets, such as Kitti [2], ETH [3], DALY [4] and UCLA HHOI [41], which are of high interest to many applications. To also justify our estimation performance quantitatively, we validate our performance on public dataset Human3.6M [5] as well, which has the ground truth of all joints on various human actions and is widely used in 3D human pose estimation and action recognition. The flowchart of our proposed system is shown in Fig. 1.

Manuscript received August 13, 2019; accepted November 2, 2019. Date of publication November 14, 2019; date of current version October 29, 2020. This article was recommended by Associate Editor T. Zhang. *(Corresponding author: Renshu Gu.)*

R. Gu and J.-N. Hwang are with the Electrical and Computer Engineering, University of Washington, Seattle, WA 98105 USA (e-mail: renshugu@uw.edu).

G. Wang and Z. Jiang are with the Electrical and Computer Engineering Department, University of Washington, Forks, WA 98331 USA.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2019.2953678

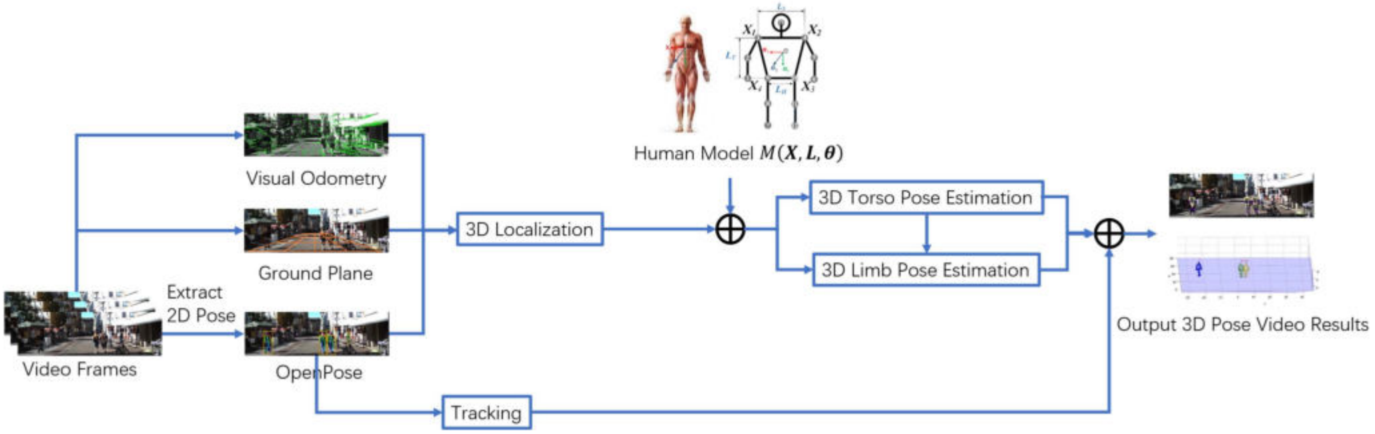


Fig. 1. Overall flowchart of the proposed algorithm.

In summary, our major contributions are: (1) we propose a pipeline that integrates visual odometry, 3D human pose estimation, and exploit the temporal information using a powerful tracking method. (2) We formulate the torso estimation as a Perspective-N-Point (PNP) problem and provide a highly efficient solution. For each limb, we formulate and solve an optimization problem. With the hierarchical problem solving structure, we greatly reduce complexity in high dimensional pose space. Moreover, each limb will not interfere with one another. (3) We design an effective occlusion handling strategy utilizing keypoint confidence in case of missing or erroneous 2D human pose estimation. (4) We provide a variety of experiments on natural videos containing multiple peoples from fixed and moving cameras, which are lacking in the current literature but critical in today's applications. Our solution provides great new opportunities to understand and predict human behaviors in natural videos.

The organization of the paper is as follows: In Section II, we review some related works of 2D/3D human pose estimation based on monocular cameras. The preprocessing including 2D tracking and 3D localization of our proposed scheme is then introduced in Section III. In Section IV, the hierarchical 3D pose estimation is discussed in details. Experimental results are presented in Section V, followed by the conclusions in Section VI.

## II. Related Work

### A. 2D Human Pose Estimation From Monocular Cameras

As a crucial task to facilitate analyses of human actions and activities in image and videos, human pose estimation in 2D based on monocular cameras has been well studied. Recent efforts on deep learning approaches, mainly CNN based [1], [6]]–[15], [45], show reliable results for multiple people. It is relatively easier to label 2D human pose without the need of experimental settings; therefore, acquiring 2D training data is less expensive. The performance on the MPII benchmark [16] has become saturated in the past three years, reaching more than 90% percentage of correct keypoints (PCKh) with a successful predicted human joint localization being within 50% of the head segment length to the ground truth joint (PCKh@0.5) [12]–[15].

### B. 3D Human Pose Estimation From Monocular Cameras

Unlike 2D human pose estimation, 3D human pose estimation from monocular cameras is far from mature. Many early approaches [17], [43], [44] are based on appearance models (e.g., silhouettes) and perform tracking using stochastic search with kinematic constraints. However, silhouette extraction can be unreliable due to complex backgrounds, occlusions, and moving cameras. When deep learning prospers, researchers first turn to 3D training data to solve the problem. 3D training data are obtained by the MoCap system in constrained environments. Later, to tackle the more challenging task of 3D pose estimation in the wild, some researchers find 3D training data not enough, and use 2D training data in addition. Related methods can be roughly grouped into 3 categories. The first two categories are called Direct3D methods as they need 3D MoCap training data. In contrast, the third category, which our method falls within, does not necessarily need 3D training data. The three categories are as follows.

1) Training with 3D Input and Ground Truth: 3D human pose estimation methods are directly trained and tested based on multi-view 3D MoCap data and corresponding ground truth of 3D joints [18], [19]. Li and Chan [18] pre-train their network with maps for 2D joint classification, and use a multi-task framework to jointly train pose regression and body part detectors. In [19] the authors propose a framework that can be interpreted as a special form of structured support vector machines where the joint feature space is discriminatively learned. This category of methods depends heavily on 3D training data, which is lacking in the current literature.
2) Joint estimation from multiple sources: jointly solves 2D and 3D pose estimation from image sequences [25], [26], [28]–[30], [42]. Methods in this category try to compensate the lack of 3D training data by incorporating 2D training data. Still, a significant amount of training data is required. The position of human subject(s) in the image sequences affects the training.
3) The 3D Pose Inference from Estimated 2D Pose: the problem is tackled with two steps: first, detect the 2D joints by a 2D pose detector [20], [21]; second, estimate 3D poses from 2D poses. In other words, methods in

this category use the 2D results as an intermediate step. More specifically, Ramakrishna *et al.* [22] represent a 3D pose by a linear combination of a set of base poses, which are learned from motion databases, by minimizing the reprojection error directly in the high dimensional pose space. To overcome this high dimensional search problem, Wang *et al.* [23] further extend the work in [22] by enforcing the length proportions of eight limbs with respect to the body length to be constant. In recent work [24], Wang *et al.* represent 3D poses by a sparse combination of bases which encode structural pose priors to reduce the lifting ambiguity. Their system outputs $K$ candidate 3D poses and improve 3D pose estimates by post-processing as well as exploiting temporal cues. In [27], Zhou *et al.* solve the correspondence between video and 3D motion capture data. Along the same line of research, we estimate 3D pose from 2D intermediate results in a hierarchical fashion, where we apply a prior human model that allows proportions of bone lengths to be adaptively determined.

While 3D human pose estimation of a single person based on monocular moving camera has been reported, there are less papers that have analyzed 3D multi-human pose estimation performance in natural videos, especially recordings from moving cameras such as car-mounted cameras, which are of high interest to autonomous driving, etc. A few papers [30] show results for in-the-wild data such as MPII and MPII-INF-HP, but they do not have 3D quantitative results on multi-person videos. This paper, on the other hand, qualitatively and quantitatively addresses 3D multi-person human pose estimation in a variety of natural videos from fixed or moving cameras, including transportation scenarios, daily activities, sports, etc. We will compare our method with several state-of-the-art 2D-to-3D methods. Our aim is not to show increased accuracy on well-trained datasets recorded in experimental settings, which can be better handled by deep learning approaches, as we believe there are certain limitations in this branch of work. Instead, we show superior performance on natural multi-person videos recorded by fixed or moving cameras.

## III. 2D Pose Tracking and 3D Localization by Moving Cameras

### A. Notations

First, as summarized in Table I, we define some notations of our pose estimation system, where $(\cdot)_t$ represents the corresponding variable at the time $t$.

### B. 2D Pose Estimation and Tracking

We take advantage of the recent advances in 2D human pose estimation using a deep neural network (DNN) based human pose detector, OpenPose [1], which independently detects 2D joints for every image frame of the video. Since the 2D pose estimation is not the main focus of this paper, we just take the 2D pose estimated by OpenPose as the input initial value of our system.

TABLE I

Notations of the System

| Symbol | Notations |
|---|---|
| $\boldsymbol{R}^{(C)}$ | Rotation matrix of the camera pose, 3 × 3 |
| $\boldsymbol{t}^{(C)}$ | Translation vector of the camera pose, 3 × 1 |
| $\boldsymbol{R}^{(H)}$ | Rotation matrix of the human pose, 3 × 3 |
| $\boldsymbol{t}^{(H)}$ | Translation vector of the human pose, 3 × 1 |
| $\boldsymbol{K}$ | Intrinsic camera matrix, 3 × 3 |
| $\boldsymbol{n}^{(G)}$ | Normal vector of the ground plane, 3 × 1 |
| $h^{(G)}$ | The camera height to the ground plane, scalar |
| $\boldsymbol{P}$ | 3D point in the world coordinates, $(X, Y, Z)$ |
| $\boldsymbol{p}$ | 2D point on the image plane, $(x, y, 1)$ |
| $\boldsymbol{X}^{(H)}$ | 3D joint in the human model coordinates, $(X, Y, Z)$ |
| $\boldsymbol{X}^{(C)}$ | 3D joint in the camera coordinate, $(X, Y, Z)$ |
| $\boldsymbol{X}^{(W)}$ | 3D joint in the world coordinates, $(X, Y, Z)$ |
| $\boldsymbol{x}$ | 2D joint on the image plane, $(x, y, 1)$ |

OpenPose only focuses on single images. However, there are several drawbacks of single frame-based pose estimation, which are listed as follows.

a) *The occlusion cannot be easily handled*. When some joints are occluded, the single frame-based pose estimation becomes unreliable. If multiple people are close to each other, joints from different people can be easily tangled together. The pose even cannot be estimated when full occlusion happens.

b) *Temporal information is not well exploited for further analysis*. Since there is no association across frames, it becomes unclear which person/joint corresponds to which person/joint across frames. Without association in the time domain, it is hard to perform further analysis, such as behavior analysis, anomaly detection and speed estimation.

To address the above drawbacks, we apply a multiple-object tracking method to solve occlusion and association problem. We adopt TrackletNet Tracking (TNT) [31] for human tracking under fixed or moving cameras since it shows great performance when handling occlusions. Based on the detection results from OpenPose, the tracklets are generated based on intersection-over-union (IOU) and appearance similarity between two adjacent frames. Then the tracklet based graph model is built with each tracklet being treated as a node in the graph. For every edge between two nodes, the connectivity is defined by a pre-trained multi-scale TrackletNet, which measures the similarity between two tracklets. Then clustering is conducted to minimize the total cost on the graph, so that the tracklets from the same ID can be merged into one group. The details of TNT can be found in [31]. Simply put, the goal of tracking is to obtain the unique ID for each detected person from OpenPose, i.e.,

$$ID_{i,t} = \mathfrak{T}\left(D_{i,t}\right), \tag{1}$$

where $D_{i,t}$ is the $i$-th detection at frame $t$ and $\mathfrak{T}(\cdot)$ is the tracker function. The output of the tracking is the unique person ID of the detection $D_{i,t}$. Since the detection results from OpenPose sometimes are very noisy, Kalman filter is applied for smoothing for each individual joint.

### *C. 3D Pose Localization by Visual Odometry and Ground Plane Estimation*

We use state-of-the-art semi-direct visual odometry (SVO) [19] technique to calculate the camera trajectory, so as to infer the 3D location of the human to be pose estimated. From SVO we can localize the camera position and pose, as well as the ground plane. Subsequently, the foot location for each human in the world coordinates can be estimated. For every 3D human-foot point $\boldsymbol{P}$ of the OpenPose detected person, who stands on the ground plane, the following two constraints should be followed,

$$\boldsymbol{K}\left(\boldsymbol{R}^{(C)}\boldsymbol{P}+\boldsymbol{t}^{(C)}\right)=s\boldsymbol{p}, \tag{2}$$

$$\boldsymbol{n}^{(G)}\boldsymbol{P}+h^{(G)}=0, \tag{3}$$

where the camera pose $\left[\boldsymbol{R}^{(C)}|\boldsymbol{t}^{(C)}\right]$ can be estimated by SVO [32], while the ground plane $\left(\boldsymbol{n}^{(G)},h^{(G)}\right)$ can be estimated by [33]. Note that, $\boldsymbol{p}$ is the 2D projection of the 3D point $\boldsymbol{P}$ on the image plane with scale factor $s$, as indicated in Eq. (2), and Eq. (3) specifies the ground plane constraint. The 3D point $\boldsymbol{P}$ can thus be localized as a function of $\boldsymbol{R}^{(C)}$, $\boldsymbol{t}^{(C)}$, $\boldsymbol{n}^{(G)}$, and $h^{(G)}$(as defined in Table I), derived from Eq. (2) and Eq. (3), i.e.,

$$\begin{aligned}&\boldsymbol{P}\left(\boldsymbol{R}^{(C)},\boldsymbol{t}^{(C)},\boldsymbol{n}^{(G)},h^{(G)}\right)\\&=\left(\boldsymbol{K}\boldsymbol{R}^{(C)}\right)^{-1}\left(\frac{\boldsymbol{n}^{(G)T}\left(\boldsymbol{K}\boldsymbol{R}^{(C)}\right)^{-1}\boldsymbol{K}\boldsymbol{t}^{(C)}-h^{(G)}}{\boldsymbol{n}^{(G)T}\left(\boldsymbol{K}\boldsymbol{R}^{(C)}\right)^{-1}\boldsymbol{p}}\boldsymbol{p}-\boldsymbol{K}\boldsymbol{t}^{(C)}\right),\end{aligned} \tag{4}$$

Then the absolute height, $H$, of the estimated human can be obtained by,

$$H=\frac{\|\boldsymbol{P}_{bl}-\boldsymbol{P}_{br}\|\,h}{w}, \tag{5}$$

where $\boldsymbol{P}_{bl}$ and $\boldsymbol{P}_{br}$ are the two bottom 3D points corresponding to the two bottom points of the detection bounding box on the ground plane, and $w$ and $h$ are the width and height of the 2D detection bounding box.

## IV. Proposed 3D Pose Estimation

### *A. Flexible Hierarchical 3D Human Body Model*

To estimate reasonable 3D poses for human, we adopt a flexible hierarchical 3D human body model. The torso is at the top level in the human body model hierarchy. The upper limbs are at the second level, which depends on the top-level pose of the torso. The lower limbs are at the third level, which depends on the second-level upper limbs, and thus on the top-level pose of the torso. The human body model is defined as $M\left(\boldsymbol{X},\boldsymbol{L},\boldsymbol{\theta}\right)$, parameterized by joints $\boldsymbol{X}$, bone lengths $\boldsymbol{L}$, and angles $\boldsymbol{\theta}$. In our definition of the human model, all the joints are flexible, which can be decomposed to different poses. However, the joints are also constrained by the bone length $\boldsymbol{L}$ and joint angle $\boldsymbol{\theta}$. In particular, there are 13 joints used in our human models. In the human model coordinate system, the origin is defined as the center of the torso plane, which also determines the 3D locations of shoulder and hip joints. An example of the flexible human model is shown in Fig. 2.

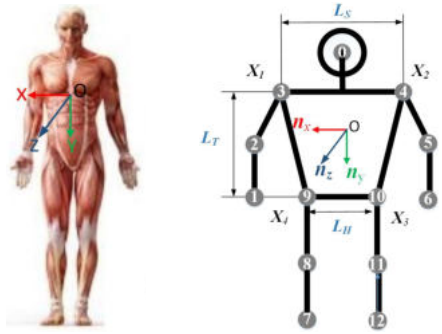


Fig. 2. Flexible hierarchical 3D human body model coordinates.

TABLE II
Angle Constraints of the Human Model as Defined in [38]

| joint | axis | $\theta^{-}$(degree) | $\theta^{+}$(degree) |
|---|---|---|---|
| shoulder | $\boldsymbol{n}_x$ | -60 | 180 |
| | $\boldsymbol{n}_y$ | -30 | 135 |
| | $\boldsymbol{n}_z$ | 0 | 180 |
| elbow | $\boldsymbol{n}_x$ | -10 | 150 |
| hip | $\boldsymbol{n}_x$ | -15 | 140 |
| | $\boldsymbol{n}_z$ | -45 | 30 |
| knee | $\boldsymbol{n}_x$ | -10 | 150 |

Then a connectivity matrix is defined as follows to measure whether two joints $\boldsymbol{X}_i$, $\boldsymbol{X}_j$ are connected in the skeleton,

$$\boldsymbol{C}\left(i,j\right)=\begin{cases}1 & \boldsymbol{X}_i,\boldsymbol{X}_j\ connected,\\0 & \text{O.W.}\end{cases} \tag{6}$$

Similarly, the matrix of bone length between each pair of joints is defined as,

$$\boldsymbol{L}\left(i,j\right)=Hl_{i,j}, \tag{7}$$

where $l_{i,j}$ is the initialized normalized bone length between connected joints $\boldsymbol{X}_i$, $\boldsymbol{X}_j$. Note that bone lengths $\boldsymbol{L}\left(i,j\right)$ are proportional to the body height. We use this set of bone lengths $l_{i,j}$ to initialize, and then include bone lengths as parameters for optimization. This way, we can obtain a fine-tuned final estimation.

Angle constraints are summarized in Table II and Fig. 3. Denote the 3D coordinate system of torso pose as $\boldsymbol{n}_x$, $\boldsymbol{n}_y$, $\boldsymbol{n}_z$, as shown in Fig. 2, and $\boldsymbol{\theta}$ as the set of angle constraints. For any $u_{i,j}\in\boldsymbol{\theta}$, $u$ is a 6D tuple, denoted as $u_{i,j}=\left(i,j,\boldsymbol{n}_{i,j,1},\boldsymbol{n}_{i,j,2},\theta_{i,j}^{-},\theta_{i,j}^{+}\right)$, where $i$, $j$ are the pair of joints from one of the seven categories as shown in Table II, $\boldsymbol{n}_{i,j,1}$ and $\boldsymbol{n}_{i,j,2}$ are two axes of angle rotation plane picked from $\boldsymbol{n}_x$, $\boldsymbol{n}_y$, $\boldsymbol{n}_z$, which are shown in each sub-plot of Fig. 3, and $\theta_{i,j}^{-},\theta_{i,j}^{+}$ are the lower and upper bounds of the angle constraints. Denote $u_{i,j}=\left(x_{i,j,1},x_{i,j,2}\right)$ as the 2D coordinate of the angle rotation plane, i.e.,

$$x_{i,j,1}=\frac{\boldsymbol{n}_{i,j,1}^{T}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)}{\|\boldsymbol{X}_i-\boldsymbol{X}_j\|}, \tag{8}$$

$$x_{i,j,2}=\frac{\boldsymbol{n}_{i,j,2}^{T}\left(\boldsymbol{X}_i-\boldsymbol{X}_j\right)}{\|\boldsymbol{X}_i-\boldsymbol{X}_j\|}. \tag{9}$$

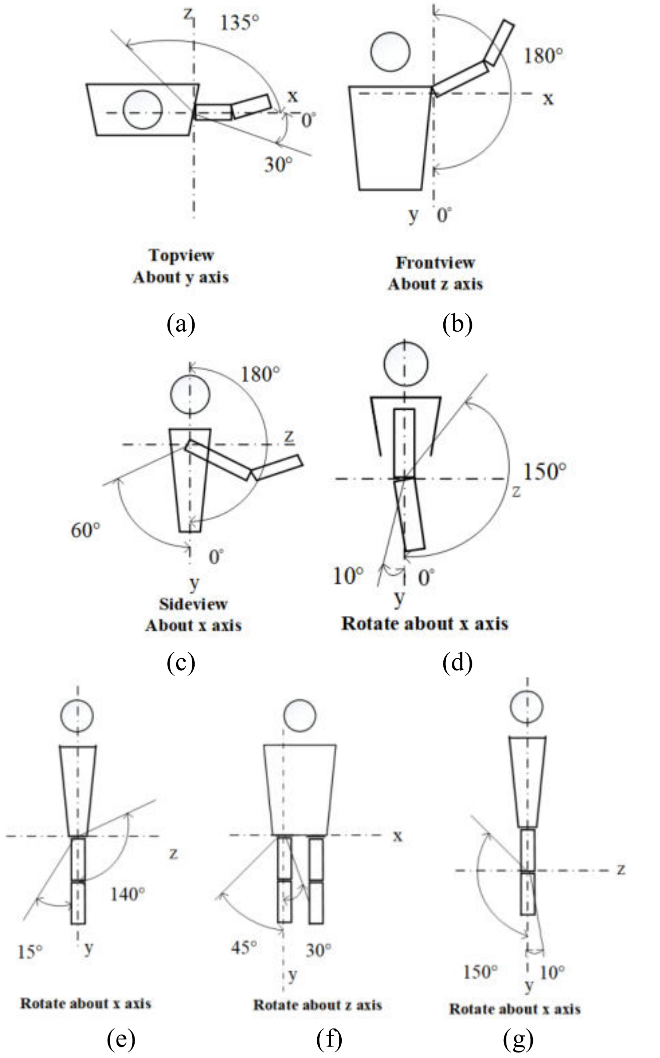


Fig. 3. Limb angle limits [38]. (a) Shoulder angle limits, top view. (b) Shoulder angle limits, front view. (c) Shoulder angle limits, side view. (d) Elbow angle limits. (e) Hip angle limits, side view. (f) Hip angle limits, front view. (g) Knee angle limits, side view.

Then the angle constraint can be represented as

$$\theta_{i,j}^{-} \leq \mathrm{Angle}\left(x_{i,j,1}, x_{i,j,2}\right) \leq \theta_{i,j}^{+}, \tag{10}$$

where $\mathrm{Angle}\,(x, y)$ is the function of $x, y$ that outputs the angle of the point $(x, y)$ on the 2D space.

### *B. Hierarchical 3D Human Pose Estimation*

Our proposed pose estimation is processed in a hierarchical way, i.e., the torso pose is estimated first, followed by the upper and lower limb pose estimation. Several advantages of this hierarchical approach are discussed below.

- **Robustness in multi-person localization and pose estimation**. As we know, the four torso joints are least flexible than the joints on the limbs, which can be treated as a rigid object, and they usually form a 3D regular plane. Based on the Perspective-N-Point (PNP) algorithm [34], the torso poses can be estimated and localized related to the camera easily and robustly. On the other hand, if we also take into account the joints on the limbs at this point, then the 3D torso pose cannot be easily and accurately inferred.
- **Simplify the estimation**. Since the limb pose is largely dependent on the torso pose, after we estimate the torso pose, the limb pose can be easily inferred. Moreover, the dissection of the problem greatly reduces the search space with increased efficiency. Methods that try to solve the full set of poses, i.e., 13 joints, suffer from computation complexity. On the contrary, we structure poses hierarchically and formulate torso and limb estimations respectively with a lower degree of freedoms (DoFs), enabling real-time processing capability. This hierarchical way can largely simplify the constraints and make the optimization efficiently.

*1) Torso Pose Estimation:* For each person, the camera pose can be inferred by solving a PNP problem with four pairs of 3D torso joints in the human model and 2D joints on the image plane. For each 3D and 2D pair $i$ of the $k$-th person, they should follow the projection constraint as follows,

$$s\boldsymbol{x}_{k,i} = \boldsymbol{K}\left(\boldsymbol{R}_{k}^{(H)}\boldsymbol{X}_{k,i}^{(H)} + \boldsymbol{t}_{k}^{(H)}\right), \tag{11}$$

where $\boldsymbol{X}_{k,i}^{(H)}$ represents the location of the $i$-th torso joint of the $k$-th person in the human model coordinates. Particularly, we use $\boldsymbol{X}_{k,i\in\{1,2,3,4\}}$ to denote the torso points, as shown in Fig. 2. For simplification, we denote $L_S$, $L_T$ and $L_H$ as the length of shoulder, torso, and hips, respectively as shown in Fig. 2. Then, the four torso points can be represented as

$$\begin{aligned} \boldsymbol{X}_{k,1} &= \left(\frac{L_S}{2}, -\frac{L_T}{2,}, 0\right), \quad \boldsymbol{X}_{k,2} = \left(-\frac{L_S}{2}, -\frac{L_T}{2}, 0\right), \\ \boldsymbol{X}_{k,3} &= \left(-\frac{L_H}{2}, \frac{L_T}{2}, 0\right), \quad \boldsymbol{X}_{k,4} = \left(\frac{L_H}{2}, \frac{L_T}{2}, 0\right). \end{aligned} \tag{12}$$

Given 4 pairs of joints $\{\boldsymbol{X}_{k,i}^{(C)}\}$, the torso pose $\left[\boldsymbol{R}_{k}^{(H)}|\boldsymbol{t}_{k}^{(H)}\right]$ of the $k$-th person in the camera coordinates can be represented by

$$\boldsymbol{X}_{k,i}^{(C)} = \boldsymbol{R}_{k}^{(H)}\boldsymbol{X}_{k,i}^{(H)} + \boldsymbol{t}_{k}^{(H)}. \tag{13}$$

We can also transform the world coordinates to the camera coordinate with the estimated camera pose by

$$\boldsymbol{X}_{k,i}^{(C)} = \boldsymbol{R}^{(C)}\boldsymbol{X}_{k,i}^{(W)} + \boldsymbol{t}^{(C)}, \tag{14}$$

Then, we can get the torso joint $\{\boldsymbol{X}_{k,i}^{(W)}\}$ in the world coordinates as

$$\begin{aligned} \boldsymbol{X}_{k,i}^{(W)} &= \left(\boldsymbol{R}^{(C)}\right)^{-1}\boldsymbol{X}_{k,i}^{(C)} - \left(\boldsymbol{R}^{(C)}\right)^{-1}\boldsymbol{t}^{(C)} \\ &= \left(\boldsymbol{R}^{(C)}\right)^{-1}\left(\boldsymbol{R}_{k}^{(H)}\boldsymbol{X}_{k,i}^{(H)} + \boldsymbol{t}_{k}^{(H)}\right) - \left(\boldsymbol{R}^{(C)}\right)^{-1}\boldsymbol{t}^{(C)}. \end{aligned} \tag{15}$$

*2) Limb Pose Estimation:* The limb pose estimation is built upon the estimated torso pose, which is one important feature of our hierarchical pose estimation. The limb pose is estimated based on the reprojection error, as well as the constraints of bone length $\boldsymbol{L}\,(i, j)$ and joint angle $\boldsymbol{\theta}$, which are defined in the previous section. Then the cost function of the joints on any limb is defined as follows,

$$\begin{aligned} f\left(\boldsymbol{X}_{k,i}^{(C)}\right) &= \sum_i c_{k,i}\left\|\boldsymbol{K}\boldsymbol{X}_{k,i}^{(C)} - s_{k,i}\boldsymbol{x}_{k,i}\right\| \\ &\quad + \rho_1 \sum_{u_{i,j}\in\boldsymbol{\theta}} d\left(\text{Angle}\left(x_{i,j,1}, x_{i,j,2}\right), R\left(\theta_{i,j}\right)\right) \\ &\quad + \rho_2 \sum_{i,j} \boldsymbol{C}(i,j)\, d\left(\left\|\boldsymbol{X}_{k,i}^{(C)} - \boldsymbol{X}_{k,j}^{(C)}\right\|, R\left(\boldsymbol{L}(i,j)\right)\right), \quad (16) \end{aligned}$$

where $c_{k,i}$ is the confidence score of the $i$-th joint from OpenPose, $\boldsymbol{X}_{k,i}^{(C)}$ is the $i$-th joint on the limb of the $k$-th person, and $u_{i,j}$ is a 6D tuple element from $\boldsymbol{\theta}$ defined in the human model. For arms (upper limbs), either elbow or wrist; for legs (lower limbs), eitherknee or foot. $\boldsymbol{x}_{k,i}$ is the corresponding 2D joint on the image plane, $s_{k,i}$ is the scale, $R\left(\theta_{i,j}\right) = \left[\theta_{i,j}^{-}, \theta_{i,j}^{+}\right]$ is the range of the joint angle, $R\left(\boldsymbol{L}(i,j)\right) = \left[\boldsymbol{L}(i,j) - \delta_l, \boldsymbol{L}(i,j) + \delta_l\right]$ is the range of the bone length, where $\delta_l$ is a small value which we set to $0.1\boldsymbol{L}(i,j)$ in our experiments. The distance function $d(x, R)$ measures the exponential cost between $x$ and $R$, i.e.,

$$d(x, R) = \exp\left(\min_{r\in R}\left(\frac{|x-r|}{\max(R) - \min(R)}\right)\right) - 1. \quad (17)$$

If $x$ lies in the range $R$, then the output of the function is 0; otherwise, the output is the minimum exponential absolute distance to the range $R$.

The cost function in Eq. (16) can be efficiently solved by the Powell's method [35]. After the optimization, the joint location in the world coordinates can be obtained by Eq. (15).

Thanks to the use of our hierarchical human model, we are able to dissect the calculation and allow parallel processing efficiently for the single frame pose initialization. Below is the flowchart of our limb parallel processing design for a single subject. For multiple persons, each person can be processed in parallel as well, which is shown in Fig. 4.

*3) Occlusion Reasoning:* We utilize keypoint confidence and geometry reasoning to handle occlusion. For the joint in the camera coordinates, if it is occluded, then the reprojection constraint needs to be relaxed since the point is hard to be seen from the camera view. In other words, we are interested in the probability that a joint is both detected and visible (not occluded), which can be formulated as

$$P(V, D) = P(V|D)\, P(D), \quad (18)$$

where $P(D)$ is the detection probability, i.e., the joint confidence $c_{k,i}$ computed by OpenPose, $P(V|D)$ is the probability of the visibility of the detected joint, and $P(V, D)$ is the probability that a joint is both detected and visible.

To approximate $P(V|D)$, a sigmoid function is adopted, i.e.,

$$P(V|D) = \frac{1}{1 + \exp(Z - Z_c)}, \quad (19)$$

where $Z$ is the depth of the target joint, and $Z_c$ is the depth of the center of the torso. If $Z$ is larger than $Z_c$, then the joint is more likely to be occluded by the torso; otherwise, it is more likely to be visible. This situation is very common, especially when the camera is on the side view of the human body. Rather than using $c_{k,i}$ as the weight of reprojection error directly, we use $P(V, D)$ as the weight. Then Eq. (16) is reformulated as

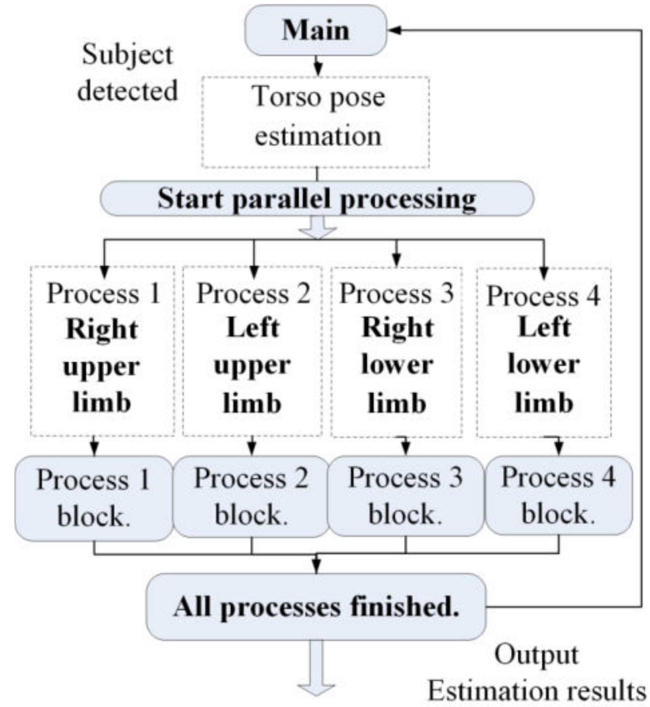

Main
Subject detected
Torso pose estimation
Start parallel processing
Process 1 Right upper limb
Process 2 Left upper limb
Process 3 Right lower limb
Process 4 Left lower limb
Process 1 block.
Process 2 block.
Process 3 block.
Process 4 block.
All processes finished.
Output Estimation results

Fig. 4. Flowchart of parallel processing.

$$\begin{aligned} f\left(\boldsymbol{X}_{k,i}^{(C)}\right) &= \sum_i P(V, D)\left\|\boldsymbol{K}\boldsymbol{X}_{k,i}^{(C)} - s_{k,i}\boldsymbol{x}_{k,i}\right\| \\ &\quad + \rho_1 \sum_{u_{i,j}\in\boldsymbol{\theta}} d\left(\text{Angle}\left(x_{i,j,1}, x_{i,j,2}\right), R\left(\theta_{i,j}\right)\right) \\ &\quad + \rho_2 \sum_{(i,j)} \boldsymbol{C}(i,j)\, d\left(\left\|\boldsymbol{X}_{k,i}^{(C)} - \boldsymbol{X}_{k,j}^{(C)}\right\|, R\left(\boldsymbol{L}(i,j)\right)\right). \quad (20) \end{aligned}$$

### *C. Joint Optimization With Temporal Constraints to Deal With Missing Poses*

After we initialize the torso pose and limb pose in a hierarchical way for each frame, we also want to take into consideration of temporal constraints to handle missing poses and occlusions with smoothness constraints. The smoothness term $\lambda_3 \sum_t^T \sum_i^N \left\|\boldsymbol{X}_{i,t}^{(W)} - \boldsymbol{X}_{i,t-1}^{(W)}\right\|^2$ in (21) penalizes the objective function if the 3D coordinates $\boldsymbol{X}_{i,t}^{(W)}$ of a joint $i$ are too far away from each other between two adjacent frames $t-1$ and $t$. Rather than estimating the joint location in the camera coordinates, we fine-tune the 3D joint location in the world coordinates directly since the joint locations are usually very smoothing in the world coordinates and independent to the camera pose. The cost function with temporal constraints

TABLE III
SUMMARY OF DIFFERENT DATASETS

| Dataset | Moving Camera | Multiple People | Ground Truth Available | Number of Actions |
|---|---|---|---|---|
| Kitti [2] | yes | yes | no | N/A |
| ETH [3] | yes | yes | no | N/A |
| DALY [4] | yes | yes | no | N/A |
| UCLA HHOI [41] | no | yes | yes | 4 |
| Human3.6M[5] | no | no | yes | 15 |
| UWHHI (ours) | yes | yes | yes | 2 |

involved is defined as follows,

$$
\begin{aligned}
f\left(\boldsymbol{X}_{i,t}^{(W)}\right) 
&= \sum_{t}^{T}\sum_{i}^{N} P\left(V,D\right)\left\|\boldsymbol{K}\left(\boldsymbol{R}_t^{(C)}\boldsymbol{X}_{i,t}^{(W)}+\boldsymbol{t}_t^{(C)}\right)-s_{i,t}\boldsymbol{x}_{i,t}\right\| \\
&\quad+\lambda_1\sum_{t}^{T}\sum_{u_{i,j}\in\boldsymbol{\theta}} d\left(\mathrm{Angle}\left(x_{i,j,1},x_{i,j,2}\right),R\left(\theta_{i,j}\right)\right) \\
&\quad+\lambda_2\sum_{t}^{T}\sum_{(i,j)}\boldsymbol{C}\left(i,j\right)d\left(\left\|\boldsymbol{X}_i^{(W)}-\boldsymbol{X}_j^{(W)}\right\|,R\left(\boldsymbol{L}\left(i,j\right)\right)\right) \\
&\quad+\lambda_3\sum_{t}^{T}\sum_{i}^{N}\left\|\boldsymbol{X}_{i,t}^{(W)}-\boldsymbol{X}_{i,t-1}^{(W)}\right\|^2, \qquad (21)
\end{aligned}
$$

Since we estimate the pose of each person individually, the person index $k$ is dropped for simplification. In addition to the cost defined by Eq. (20), a temporal smoothness constraint is added in the last term of Eq. (21).

## V. Experiments and Analysis

We test our method on several videos, such as KITTI dataset [2], ETH [3] and videos in DALY dataset [4] for qualitative evaluations. We also test videos in UWHHI and Human3.6M [5], which have corresponding 3D joint ground truth, for quantitative evaluations. A summary of all the datasets we tested on are listed in Table III. Moreover, ablation study is conducted on different parameter settings. All of our programs are run on Windows 10, using an Intel Core i5-6300HQ CPU@2.30Ghz, 2301MHz, 4 Core Processor. Some qualitative 3D pose estimation results can be found in the following link: https://youtu.be/YgQ0pF57mSU.

### A. Qualitative Evaluation

*1) KITTI Dataset:* KITTI datasets are captured by driving around the mid-size city of Karlsruhe, in rural areas and on highways. It is a highly popular dataset for autonomous driving research [2]. We choose some sequences from this dataset to show our capability to estimate pedestrians' 3D poses using car-mounted monocular camera. Figure 5 shows a snapshot of 3D pose estimation of pedestrians and cyclers based on our proposed scheme.

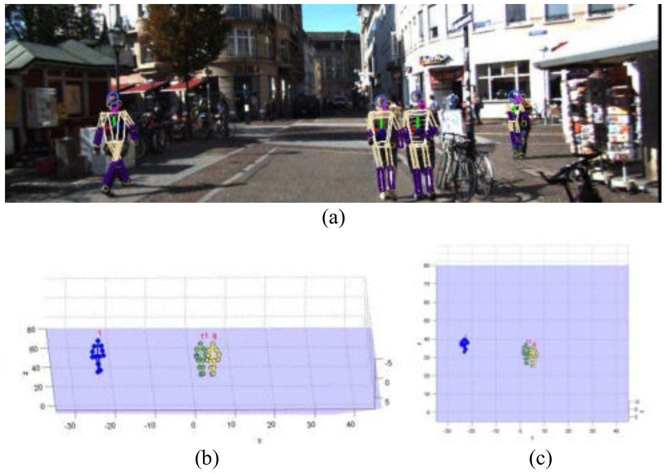


Fig. 5. 3D human pose estimation for multiple people on the street - kitti_2011_09_29. (a) Street video and human model reprojection. (b) Estimated 3D poses, front view. (c) Estimated 3D poses, top view.

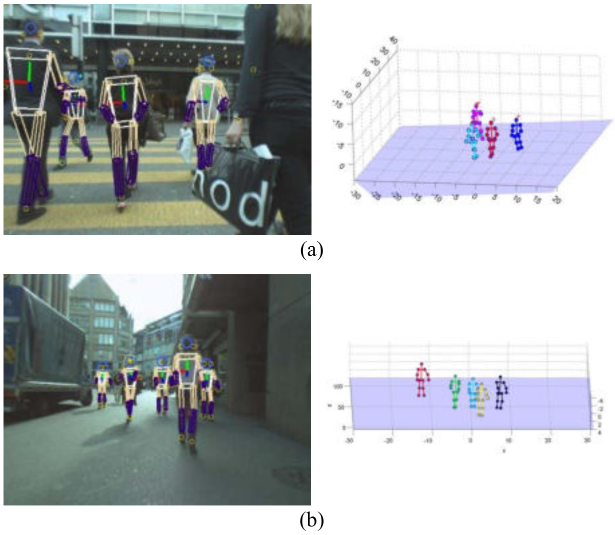


Fig. 6. (a) 3D human pose estimation for multiple people on the street – ETH dataset CROSSING sequence. (b) 3D human pose estimation for multiple people on the street – ETH dataset LINTHESCHER sequence.

*2) ETH Dataset:* ETH dataset is a dataset designed for challenging tasks of multi-person tracking. ETH dataset features transportation scenarios that contain dense pedestrians [3]. Data are recorded using a pair of AVT Marlins F033C mounted on a chariot, with a resolution of 640 × 480, and a frame rate of 13-14 FPS. We use only the left camera sequence. Our experiments show that we can also effectively perform multi-person 3D pose estimation.

*3) DALY Dataset:* DALY dataset is a dataset consists of daily activities. We find result of our method looks more natural and smoother compared to [24]. Furthermore, thanks to the powerful 2D pose deep learning OpenPose predictions, the method can handle occlusion to a certain extent. Figure 7 shows screenshots of results for this video. Figure 7(b) and 7(c) show examples of successful 3D estimation with occluded hand and/or elbow.

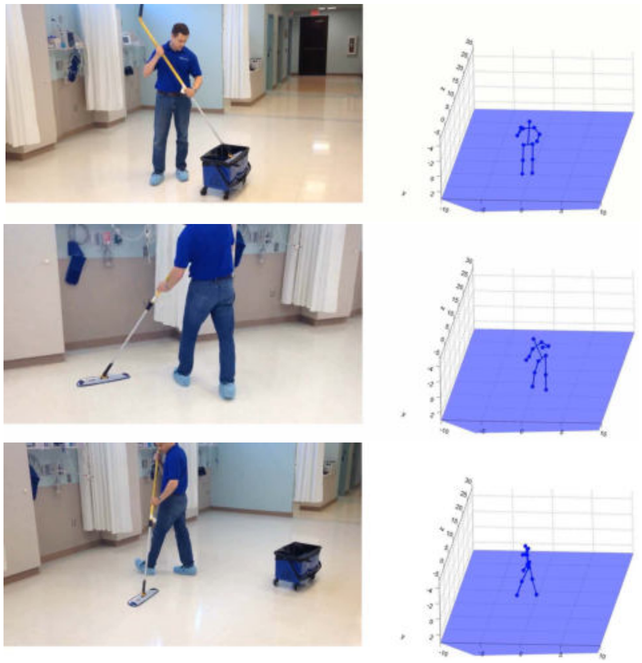


Fig. 7. Screenshots of DALY dataset "mop ground".

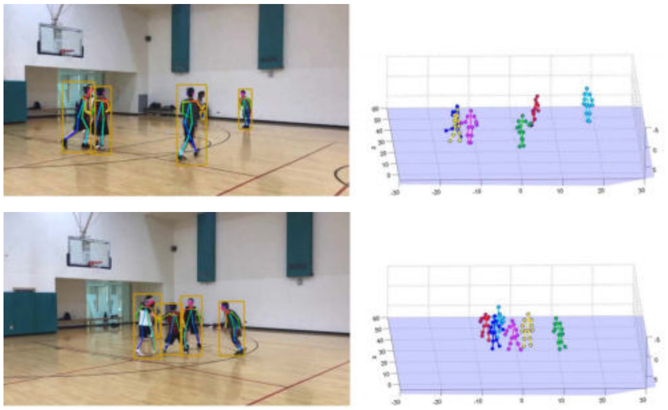
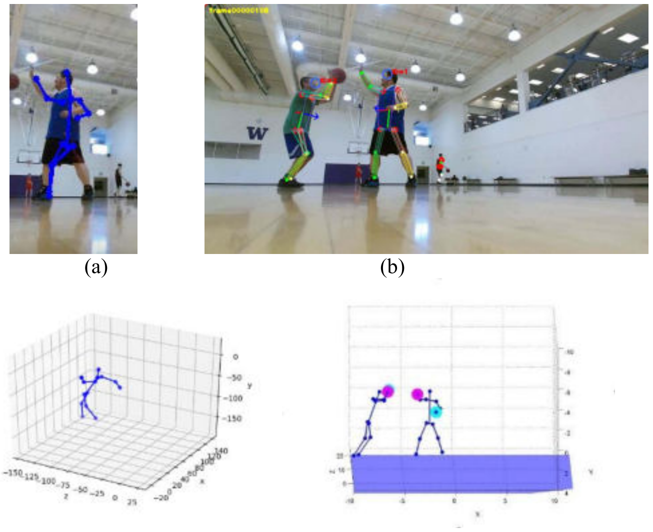


Fig. 8. Multi-person 3D human pose estimation for basketball scenario.



(a) (b)

(c) (d)

Fig. 9. Example of our method vs. hg3d on UWHHI "basketball". (a) 2D estimates of hg3d. (b) 2D reprojection of our method. (c) 3D estimates of hg3d. (d) 3D estimates of our method.

TABLE IV

DIFFERENCES COMPARED TO COMPETING METHOD

| Method | Single frame/ sequence | Multiple People | Requires cropping/ No need to crop |
|---|---|---|---|
| hg3d | Single frame | no | Requires cropping |
| Ours | sequence | yes | No need to crop |

TABLE V

UWHHI. AVERAGE 3D JOINT ERRORS IN mm

| Video | Person | Ours | SMPLify [20] | Hg3d [32] |
|---|---|---|---|---|
| Shake Hands | S0 (green hoody) | **76.5** | 116.0 | 128.9 |
| | S1 (yellow hoody) | **92.1** | 143.1 | 162.9 |
| | S2 (white shirt) | **62.6** | 140.1 | N/A |
| Basketball | S0 (black tank) | **112.2** | 162.9 | 127.6 |

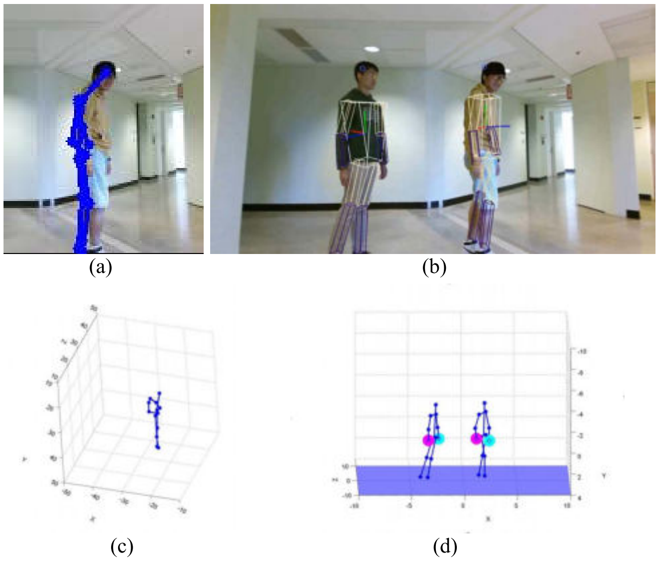
*4) Basketball Scenario:* We recorded some real-world basketball scenario videos, and test our method on these videos. Even under severe occlusion and fast motion, our method can reconstruct the 3D poses for multiple people in such scenario.

## *B. Quantitative Evaluation*

*1) UWHHI:* We also record multi-person moving camera data with ground truth, using Kinect One, as such dataset is lacking in the literature. Horizontal resolution of Kinect One is 0.75 mm per pixel at 0.5 m distance and 3 mm per pixel at 2 m distance. Depth resolution is about 1.5 mm at 0.5 m, and 3 mm at 3 m. We refer to it as University of Washington Human Human Interaction (UWHHI) data. To the best of our knowledge, none of the state-of-the art methods report quantitative evaluation on multi-person using a monocular moving camera. Figure 9 and Figure 10 show comparisons to state-of-the-art deep learning method hg3d [30]. As shown in Table III, our method outperforms hg3d for multi-person human pose estimation using a moving camera in natural scenarios.

The advantages of our method over competing method hg3d is listed in Table V. hg3d processes single frame, and only handles single person so it needs the person to be roughly at the center of the image. Therefore, in a natural video, when the person is not at the center of the image, cropping is required. And as shown in Figure 9 and Figure 10, even after cropping, hg3d can give unreasonable poses. Please see Figure 9 (c) in 3D corresponding to Figure 9 (a) in 2D, and Figure 10 (c) in 3D corresponding to Figure 10 (a) in 2D. On the other hand, our method always give reasonable pose thanks to our constraints.

*2) UCLA HHOI:* We also test our 3D pose estimation performance on UCLA Human Human Interaction (HHOI)



(a) (b)

(c) (d)

Fig. 10. Example of our method vs. hg3d on UWHHI "shake hands". (a) 2D estimates of hg3d. (b) 2D reprojection of our method. (c) 3D estimates of hg3d. (d) 3D estimates of our method.

TABLE VI

UCLA HHOI. Average 3D Joint Errors in mm. ∗ Indicates Results Which Are Obtained From the Original Paper

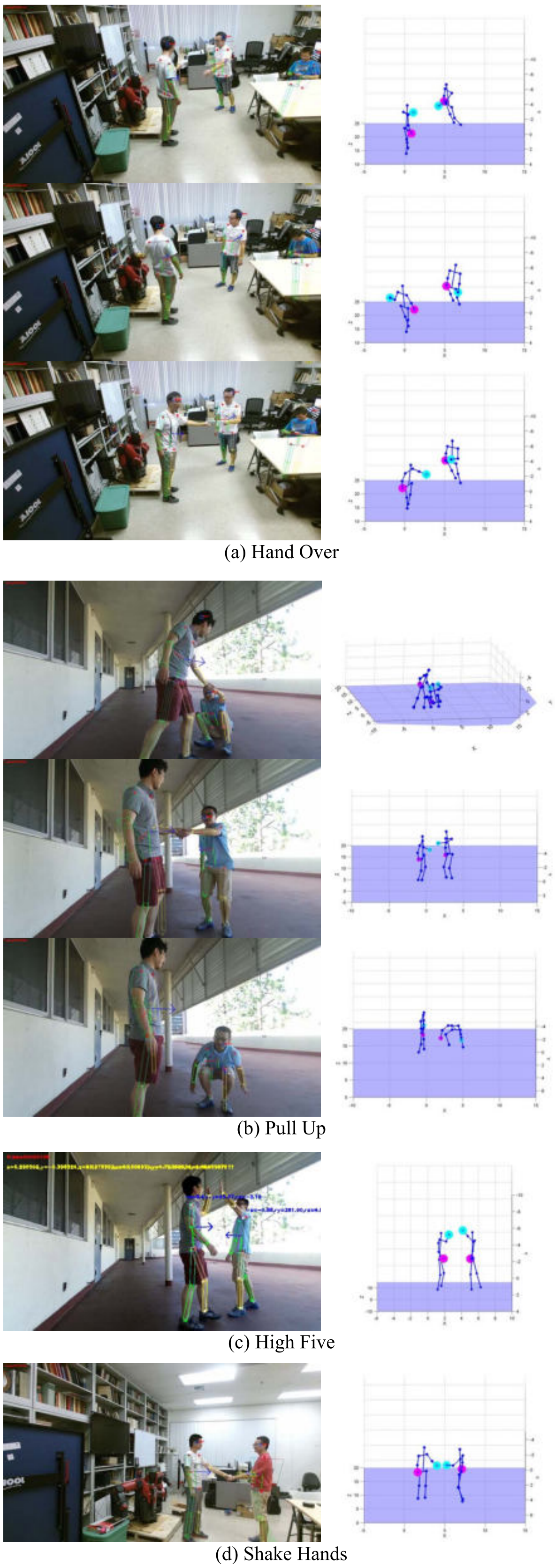
| Video | Ours | SMPLify [20] | Xiao* [40] (s+p) | Xiao* [40] (p) | Xiao* [40] (s) |
|---|---|---|---|---|---|
| Hand Over | **88.8** | 136.0 | 101.9 | 102.5 | 105.2 |
| Pull Up | **113.9** | 154.6 | 124.8 | 132.4 | 139.8 |
| Shake Hands | **77.4** | 122.1 | 118.6 | 129.0 | 113.1 |
| High Five | **94.2** | 135.6 | 96.1 | 103.0 | 98.4 |

dataset [41]. Table VI shows comparison with SMPLify [20] and Xiao's methods [40]. Here, $s$ stands for skeleton-LSTM and $p$ stands for patch-LSTM. Our method outperforms the others on this dataset by a large margin.

Note that our method is not directly compared with the deep learning results of this dataset reported in [41], because [41] uses the same dataset for training, while our method doesn't. None of the methods listed in Table VI uses training data in UCLA HHOI dataset.

*3) Human3.6M:* Our pose estimation algorithm is targeted at the challenging moving camera scenarios in uncontrolled environments. However, to the best of our knowledge, there is no public dataset of such kind with 3D ground truth available. Therefore, we validate our method on Human3.6M, which is recorded by static cameras. The Human3.6M dataset contains 3.6M human poses from actors. The videos are captured in a controlled environment from multiple calibrated static cameras and accurate 3D poses are measured using a MoCap system.

We validated our results on Human3.6M using the protocol as [20], where frames from subjects S9 and S11 are used for testing. Table VII shows that our method outperforms state-of-the-art (2018) method [24] on almost all of the actions listed, where the best performance is shown in black bold font, while the second best is shown in blue bold font. Our



(a) Hand Over

(b) Pull Up

(c) High Five

(d) Shake Hands

Fig. 11. UCLA HHOI dataset results.

method achieves quite rivaling performance as SMPLify [20], despite that SMPLify trains a regressor from the SMPL body shape to the 3D joint representation used in the dataset, while

TABLE VII
QUANTITATIVE RESULTS ON HUMAN3.6M. ERRORS ARE IN mm

| Method (2D to 3D) | Pose Class | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Directions | Discussion | Greeting | Phoning | Photo | Posing | Purchases | Sit | Waiting | Walk | Walk Together |
| Akhter & Black[36] | 199.2 | 177.6 | 197.8 | 176.2 | 186.5 | 195.4 | 167.3 | 160.7 | 181.9 | 198.6 | 192.7 |
| Ramakrishna[22] | 137.4 | 149.3 | 154.3 | 157.7 | 158.9 | 141.8 | 158.1 | 168.6 | 161.7 | 174.8 | 150.2 |
| Zhou[37] | 99.7 | 95.8 | 116.8 | 108.3 | 107.3 | 93.5 | 95.3 | 109.1 | 102.2 | 110.4 | 115.2 |
| SMPLify[20] | **62.0** | **60.2** | **76.5** | **92.1** | **77.0** | **73.0** | **75.3** | **100.3** | **77.3** | 86.8 | **81.7** |
| Wang[24] | 90.3 | 117.6 | 111.0 | 123.5 | 154.9 | 100.5 | 97.3 | 130.6 | 110.3 | **65.0** | 88.0 |
| Ours | **75.9** | 94.6 | **75.2** | **106** | **99** | **72.3** | **94.4** | **106** | **76.3** | **75.3** | **78.6** |

TABLE VIII
ABLATION STUDY. HUMAN3.6M. ERROR IN mm

| | Directions | Discussion | Greeting | Phoning | Photo | Posing | Purchases | Sit | Waiting | Walk | Walk T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline (No constraint) | 111.8 | 162.5 | 121.8 | 117.8 | 125.8 | 119.2 | 156.5 | 144.4 | 137.9 | 113.9 | 113.6 |
| Baseline+BLC+TC (W/O angle constraint) | 83.4 | 91.9 | 78.8 | 102.5 | 95.9 | 75.9 | 110.1 | 112.9 | 76.3 | 84.7 | 85.4 |
| Baseline+AC+TC (W/O bone length constraint) | 101.33 | 151.5 | 109.3 | 99.8 | 104.7 | 104.2 | 148.3 | 112.6 | 118.4 | 101.3 | 105.1 |
| Baseline+AC+BLC (W/O temporal constraint) | 77.5 | 90.4 | 77.2 | 96.9 | 95 | 75 | 96.8 | 104.1 | 76.1 | 80.3 | 78.9 |
| Baseline+AC+BLC+TC | **73.9** | **86.6** | **73.8** | **95.4** | **92.7** | **71** | **93.8** | **103.4** | **71.7** | **75.5** | **74.7** |

TABLE IX
IMPACT OF VARYING PARAMETERS ON HUMAN3.6M. ERRORS IN mm

| | Directions | Discussion | Greeting | Phoning | Photo | Posing | Purchases | Sit | Waiting | Walk | Walk T |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Non Hierarchical | 79.6 | 98.6 | 83.4 | 107.9 | 109.2 | 74.7 | 103.3 | 119.4 | 83.9 | 91.5 | 91.2 |
| W/O occlusion reasoning | 77.5 | 90.4 | 77.2 | 96.9 | 95 | 75 | 96.8 | 104.1 | 76.1 | 80.3 | 78.9 |
| Hierarchical +occlusion reasoning | **73.9** | **86.6** | **73.8** | **95.4** | **92.7** | **71** | **93.8** | **103.4** | **71.7** | **75.5** | **74.7** |

we do not use any training data. Moreover, SMPLify does not consider multiple people and will not be able to handle occlusion caused by multiple people. SMPLify only considers single frame instead of video sequences. Besides that, due to the heavy fitting process of SMPLify, optimization for a single image takes about 1 minute on a common desktop CPU, while our method is significantly faster, i.e., our proposed method can run at 30 fps on an i5 laptop. More importantly, the performance on UCLA HHOI dataset suggests that SMPLify does not generalize well on natural videos.

Overall, our method performs best among the 2D to 3D methods on 'walk together', 'Posing', 'Waiting', 'Greeting' and 'Posing'. Our method shows lower accuracy than [24] on 'walking' action possibly because [24] model poses as a set of bases which is periodic. On contrary, we decide not to include any constraints of periodicity, and our method would be more generalized on data that is non-periodic.

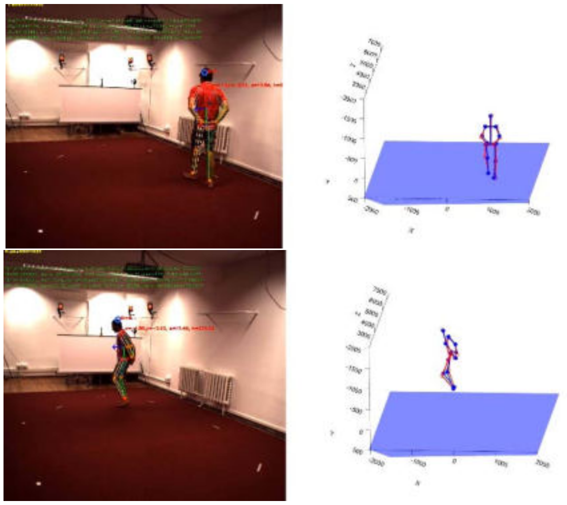


Fig. 12. Examples of 3D human pose estimation on Human3.6M. Red skeletons are ground truth from MoCap. Blue skeletons are estimated poses.

### C. Ablation Study

*1) Hierarchical vs. Non-Hierarchical:* We also investigate the impact of the hierarchical design of our method. As a control group, we disable the hierarchical estimation and optimize for 13 joints all at once. We experimented on subset of Human3.6M dataset. The results are shown in Table IX. We also experimented on HHOI dataset. The results are shown in Table XI. Our hierarchical method outperforms the non-hierarchical version. Moreover, it drastically increases computation efficiency. This advantage makes our hierarchical method promising for real-time applications.

TABLE X
ABLATION STUDY. UCLA HHOI DATASET. ERRORS IN mm

| | Hand Over | High Five | Pull Up | Shake Hands |
|---|---|---|---|---|
| Baseline (No constraint) | 101.5 | 97.6 | 139.3 | 99.7 |
| Baseline+BLC+TC (W/O angle constraint) | 91.9 | 96.8 | 119.2 | 83 |
| Baseline+AC+TC (W/O bone length constraint) | 95.5 | 105.7 | 126.3 | 83 |
| Baseline+AC+BLC (W/O temporal constraint) | 89.7 | 98 | 117.6 | 89.9 |
| Baseline+AC+BLC+TC | **88.8** | **94.2** | **113.9** | **77.4** |

TABLE XI
ABLATION STUDY. UCLA HHOI DATASET. ERRORS IN mm

| | Hand Over | High Five | Pull Up | Shake Hands |
|---|---|---|---|---|
| Non hierarchical | 95.1 | 102.9 | 119.9 | 87.1 |
| W/O occlusion reasoning | 92.3 | 101.6 | 118 | 80.3 |
| Hierarchical +Occlusion Reasoning | **88.8** | **94.2** | **113.9** | **77.4** |

*2) Varying the Parameters:* We investigate the impact of varying the parameters in Eq. (21). We disable the second term, the third term and the fourth term $\lambda_1$, $\lambda_2$ and $\lambda_3$, which correspond to angle constraint (AC), bone length constraint (BLC) and temporal constraint (TC) respectively. The results on Human3.6M are summarized in Table VIII, and the results on HHOI are summarized in Table X.

*3) Occlusion Handling:* We also show in Table IX (Human3.6M) and Table XI (HHOI) the results of w and w/o the occlusion handling strategy in Section IV.B 3). The occlusion handling strategy effectively increased accuracy by a large margin.

## VI. CONCLUSION

This paper introduces a hierarchical 3D human pose estimation method using a monocular camera on the street and in the wild. It utilizes the recent deep learning advances in 2D body joints predictions as an intermediate step, associate individuals across frames to exploit each individual's temporal information. With a human body prior, we formulate the 3D human pose estimation problem hierarchically and efficiently solve the problem in a hierarchical fashion. We first formulate the torso estimation as a PNP problem and provide a highly efficient solution. Then we dissect pose estimation for each limb, formulate and solve an optimization problem such that each limb rests in a low dimensional pose space and does not interfere with each other. Experiments show that our method qualitatively achieves natural 3D pose reconstruction results in real world videos. We also quantitatively validate our results on several action poses in a well-received dataset recorded in a constrained environment, and show it outperforms several state-of-the-art methods. Experimental results show that even though existing methods report good performance on dataset like Human3.6M, their performance may degrade on natural videos, which have much more viewpoint change, pose variations, not to mention occlusions. This suggests that existing deep learning based methods over-fit to training data and bear poor generalization capabilities. Our efficient solution to address the challenge provides great new opportunities to understand and predict human behaviors in natural videos.

## REFERENCES

[1] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7291–7299.

[2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[3] A. Ess, B. Leibe, K. Schindler, and L. Van Gool, "Robust multiperson tracking from a mobile platform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1831–1846, Oct. 2009.

[4] P. Weinzaepfel, X. Martin, and C. Schmid, "Human action localization with sparse spatial supervision," May 2016, *arXiv:1605.05197*. [Online]. Available: https://arxiv.org/abs/1605.05197

[5] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6 m: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

[6] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4724–4732.

[7] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1736–1744.

[8] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2329–2336.

[9] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[10] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[11] X. Fan, K. Zheng, Y. Lin, and S. Wang, "Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1347–1355.

[12] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 717–732.

[13] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[14] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," Feb. 2017, *arXiv: 1702.07432*. [Online]. Available: https://arxiv.org/abs/1702.07432

[15] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," 2017, *arXiv:1705.00389*. [Online]. Available: https://arxiv.org/abs/1705.00389

[16] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3686–3693.

[17] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2000, pp. 702–718.

[18] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 332–347.

[19] S. Li, W. Zhang, and A. B. Chan, "Maximum-margin structured learning with deep networks for 3D human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 2848–2856.

[20] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 561–578.

[21] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer, "Single image 3D human pose estimation from noisy observations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2673–2680.

[22] V. Ramakrishna, T. Kanade, and Y. Sheikh, "Reconstructing 3D human pose from 2D image landmarks," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 573–586.

[23] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao, "Robust estimation of 3D human poses from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2361–2368.

[24] C. Wang, Y. Wang, Z. Lin, and A. Yuille, "Robust 3D human pose estimation from single images or video sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 5, pp. 1227–1241, May 2018.

[25] H. Yasin, U. Iqbal, B. Kruger, A. Weber, and J. Gall, "A dual-source approach for 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4948–4956.

[26] E. Simo-Serra, A. Quattoni, C. Torras, and F. Moreno-Noguer, "A joint model for 2D and 3D pose estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3634–3641.

[27] F. Zhou and F. De la Torre, "Spatio-temporal matching for human detection in video," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 62–77.

[28] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net: Localization-classification-regression for human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1216–1224.

[29] B. Wandt, H. Ackermann, and B. Rosenhahn, "3D reconstruction of human motion from monocular image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1505–1516, Aug. 2016.

[30] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei, "Towards 3D human pose estimation in the wild: A weakly-supervised approach," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 398–407.

[31] G. Wang, Y. Wang, H. Zhang, R. Gu, and J.-N. Hwang, "Exploit the connectivity: Multi-object tracking with trackletnet," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 482–490.

[32] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, Apr. 2017.

[33] K.-H. Lee and J.-N. Hwang, "On-road pedestrian tracking across multiple driving recorders," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1429–1438, Sep. 2015.

[34] X.-S. Gao, X.-R. Hou, J. Tang, and H.-F. Cheng, "Complete solution classification for the perspective-three-point problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 930–943, Aug. 2003.

[35] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.*, vol. 7, no. 2, pp. 155–162, 1964.

[36] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1446–1455.

[37] X. Zhou, M. Zhu, S. Leonardos, and K. Daniilidis, "Sparse representation for 3D shape estimation: A convex relaxation approach," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1648–1661, 2016.

[38] N. Hamilton, *Kinesiology: Scientific Basis of Human Motion*, 12th ed. New York, NY, USA: McGraw-Hill, 2011.

[39] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3D human pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7025–7034.

[40] N. B. Xiaohan, P. Wei, and S.-C. Zhu, "Monocular 3D human pose estimation by predicting depth on joints," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3467–3475.

[41] T. Shu, M. S. Ryoo, and S.-C. Zhu, "Learning social affordance for human-robot interaction," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 3454–3461.

[42] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

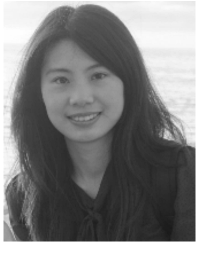
[43] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, Jan. 2005.

[44] S.-R. Ke, L. Zhu, J.-N. Hwang, H.-I. Pai, K.-M. Lan, and C.-P. Liao, "Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming," in *Proc. 7th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Aug./Sep. 2010, pp. 489–496.

[45] X. Nie, J. Feng, J. Xing, S. Xiao, and S. Yan, "Hierarchical contextual refinement networks for human pose estimation," *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 924–936, Feb. 2019.

[46] S. G. Kong and R. O. Mbouna, "Head pose estimation from a 2D face image using 3D face morphing with depth parameters," *IEEE Trans. Image Process.*, vol. 24, no. 6, pp. 1801–1808, Jun. 2015.

[47] C. Dhiman and D. K. Vishwakarma, "A robust framework for abnormal human action recognition using *R*-transform and Zernike moments in depth videos," *IEEE Sensors J.*, vol. 19, no. 13, pp. 5195–5203, Jul. 2019.



**Renshu Gu** received the bachelor's degree in electrical engineering from the University of Nanjing, China, in 2011. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Washington. Her passions include image/video analytics, computer vision, and multimedia.

**Gaoang Wang** received the B.S. degree from the Department of Electrical Engineering, Fudan University, in 2013, the M.S. degree from the Department of Electrical and Computer Engineering, University of Wisconsin-Madison in 2015, and the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Washington, in 2019. His current research interests include computer vision, machine learning, and video/image processing.

**Zhongyu Jiang** received the B.E. degree in computer science and technology from Tsinghua University, Beijing, China, in 2018. He is currently pursuing the M.Sc. degree in computer science and system with the University of Washington Tacoma.

**Jenq-Neng Hwang** (S'82–M'84–SM'96–F'01) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1981 and 1983, respectively, and the Ph.D. degree from the University of Southern California.

In summer 1989, he joined the Department of Electrical Engineering, University of Washington, Seattle, where he has been promoted to a Full Professor since 1999. He served as the Associate Chair for Research from 2003 to 2005 and 2011 to 2015. He is currently the Associate Chair of Global Affairs and International Development with the EE Department. He has written more than 330 journals, conference papers, and book chapters in the areas of machine learning, multimedia signal processing, computer vision, and multimedia system integration and networking, including an authored textbook on *Multimedia Networking: from Theory to Practice*, (Cambridge University Press). He has close working relationship with the industry on multimedia signal processing and multimedia networking. He is also a Founding Member of Multimedia Signal Processing Technical Committee of the IEEE Signal Processing Society. He is also a member of Multimedia Technical Committee (MMTC) of the IEEE Communication Society and the Multimedia Signal Processing Technical Committee (MMSP TC) of the IEEE Signal Processing Society. He was the Society's Representative of the IEEE Neural Network Council from 1996 to 2000. He received the 1995 IEEE Signal Processing Society's Best Journal Paper Award. He served as an Associate Editor for the IEEE TSP, TNN and TCSVT, TIP, and *Signal Processing Magazine* (SPM). He is also on the Editorial Board of *ZTE Communications*, ETRI, IJDMB, and JSPS journals. He has served as the Program Co-Chair of ICASSP 1998, ISCAS 2009, and the IEEE ICME 2016.