Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction

Conference Paper · June 2015

DOI: 10.1109/CVPR.2015.7298751

2 authors:

Ijaz Akhter
Australian National University
12 PUBLICATIONS **795** CITATIONS
SEE PROFILE

Michael J Black
Max Planck Institute for Intelligent Systems
377 PUBLICATIONS **30,191** CITATIONS
SEE PROFILE

Pose-Conditioned Joint Angle Limits for 3D Human Pose Reconstruction

Ijaz Akhter and Michael J. Black
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{ijaz.akhter, black}@tuebingen.mpg.de

Abstract

Estimating 3D human pose from 2D joint locations is central to the analysis of people in images and video. To address the fact that the problem is inherently ill posed, many methods impose a prior over human poses. Unfortunately these priors admit invalid poses because they do not model how joint-limits vary with pose. Here we make two key contributions. First, we collect a motion capture dataset that explores a wide range of human poses. From this we learn a pose-dependent model of joint limits that forms our prior. Both dataset and prior are available for research purposes. Second, we define a general parametrization of body pose and a new, multi-stage, method to estimate 3D pose from 2D joint locations using an over-complete dictionary of poses. Our method shows good generalization while avoiding impossible poses. We quantitatively compare our method with recent work and show state-of-the-art results on 2D to 3D pose estimation using the CMU mocap dataset. We also show superior results using manual annotations on real images and automatic detections on the Leeds sports pose dataset.

1. Introduction

Accurate modeling of priors over 3D human pose is fundamental to many problems in computer vision. Most previous priors are either not general enough for the diverse nature of human poses or not restrictive enough to avoid invalid 3D poses. We propose a physically-motivated prior that only allows anthropometrically valid poses and restricts the ones that are invalid.

One can use joint-angle limits to evaluate whether two connected bones are valid or not. However, it is established in biomechanics that there are dependencies in joint-angle limits between certain pair of bones [12, 17]. For example how much one can flex one's arm depends on whether it is in front of, or behind, the back. Medical textbooks only provide joint-angle limits in a few positions [2, 26] and the complete configuration of pose-dependent joint-angle limits for the full body is unknown.

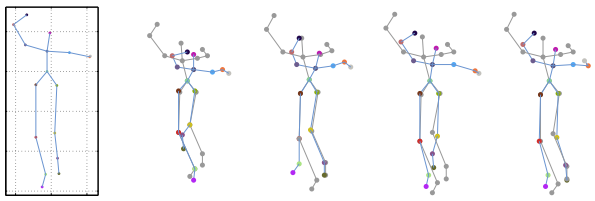
Figure 1. **Joint-limit dataset.** We captured a new dataset for learning pose-dependent joint angle limits. This includes an extensive variety of stretching poses. A few sample images are shown here.

We found that existing mocap datasets (like the CMU dataset) are insufficient to learn true joint angle limits, in particular limits that are pose dependent. Therefore we captured a new dataset of human motions that includes an extensive variety of stretching poses performed by trained athletes and gymnasts (see Fig. 1). We learn *pose-dependent joint angle limits* from this data and propose a novel prior based on these limits.

The proposed prior can be used for problems where estimating 3D human pose is ambiguous. Our pose parametrization is particularly simple and general in that the 3D pose of the kinematic skeleton is defined by the two endpoints of each bone in Cartesian coordinates. Constraining a 3D pose to remain valid during an optimization simply requires the addition of our penalty term in the objective function. We also show that our prior can be combined with a sparse representation of poses, selected from an overcomplete dictionary, to define a general yet accurate parametrization of human pose.

We use our prior to estimate 3D human pose from 2D joint locations. Figure 2 demonstrates the main difficulty in this problem. Given a single view in Fig. 2(a), the 3D pose is ambiguous [27] and there exist several plausible 3D poses as shown in Fig. 2(b), all resulting in the same 2D observations. Thus no generic prior information about static body pose is sufficient to guarantee a single correct 3D pose. Here we seek the most probable, valid, human pose.

We show that a critical step for 3D pose estimation given



(a) 2D Frame (b) 3D Pose Interpretations

Figure 2. Given only 2D joint locations in (a), there are several valid 3D pose interpretations resulting in the same image observation. Some of them are shown as colored points in (b), while the gray points represent the ground truth. Here we display the pose from a different 3D view so that the difference is clear, but all these poses project to exactly the same 2D observations.

2D point locations is the estimation of camera parameters. Given the diversity of human poses, incorrect camera parameters can lead to an incorrect pose estimate. To solve this problem we propose a grouping of body parts, called the "extended-torso," consisting of the torso, head, and upper-legs. Exploiting the fact that the pose variations for the extended-torso are fewer than for the full-body, we estimate its 3D pose and the corresponding camera parameters more easily. The estimated camera parameters are then used for full-body pose estimation. The proposed multi-step solution gives substantially improved results over previous methods.

We evaluate 3D pose estimation from 2D for a wide range of poses and camera views using activities from the CMU motion capture dataset[1]. These are more complex and varied than the data used by previous methods and we show that previous methods have trouble in this case. We also report superior results on manual annotations and automatic part-based detections [16] on the Leeds sports pose dataset. The data used for evaluation and all software is available for other researchers to compare with our results [1].

2. Related Work

The literature on modeling human pose priors and the estimation of 3D pose from points, images, video, depth data, etc. is extensive. Most previous methods for modeling human pose assume fixed joint angle limits [7, 24, 28]. Herda et al. [14] model dependencies of joint angle limits on pose for the elbow and shoulder joint. Their model cannot be used for our 2D to 3D estimation problem because it requires the unobserved rotation around the bone axis to be known. Hauberg et al. [13] suggest modeling such priors in terms of a distribution over the endpoints of the bones in the space of joint angles. We go a step further to define our model entirely on the 3D bone locations.

[1]The CMU data was obtained from http://mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

There are a number of papers on 3D human pose estimation from 2D points observed in a static camera. All such methods must resolve the inherent ambiguities by using additional information. Methods vary in how this is done. Lee and Chen [18] recover pose by pruning a binary interpretation tree representing all possible body configurations. Taylor [29] resolves the depth ambiguity using manual intervention. Barròn and Kakadiaris [4] use joint angle limit constraints to resolve this ambiguity. Parameswaran and Chellappa [21] use 3D model-based invariants to recover the joint angle configuration. BenAbdelkader and Yacoob [5] estimate limb lengths by exploiting statistical limits on their ratios. Guan et al. [11] use a database of body measurements and the known gender and height of a person to predict bone lengths. Bourdev and Malik [6] estimate pose from key points followed by manual adjustment. Jiang [15] uses Taylor's method and proposes an exemplar-based approach to prune the hypotheses. Ramakrishna et al. [23] propose an over-complete dictionary of actions to estimate 3D pose. These methods do not impose joint angle limits and can potentially estimate an invalid 3D pose.

Some of the ambiguities in monocular pose estimation are resolved by having a sequence (but not always). Wei and Chai [31] and Valmadre and Lucey [30] estimate 3D pose from multiple images and exploit joint angle limits. To apply joint angle limits, one must first have a kinematic tree structure in which the coordinate axes are clearly defined. Given only two points per bone, this is itself a seriously ill-posed problem requiring prior knowledge. Valmadre and Lucey require manual resolution to fix this issue. Our body representation simplifies this problem since it does not represent unobserved rotations about the limbs. We believe ours is the first work to propose joint-angle limits for a kinematic skeleton in Cartesian coordinates, where only two points per bone are known.

In Computer Graphics, there also exist methods for human pose animation from manual 2D annotations. Grochow et al. [10] proposed a scaled Gaussian latent variable model as a 3D pose prior. Space complexity of their method is a quadratic in the size of the training data. Wei and Chai [32] and Lin et al. [19] require additional constraints, like the distance between joints or the ground plane to be known, to resolve ambiguity in pose estimation. Yoo et al. [33] and Choi et al. [8] propose a sketching interface for 3D pose estimation. Their methods only works for the poses present in the training data.

Discriminative approaches also exist in the literature that do not require 2D point correspondence and directly estimate human pose from 2D image measurements [3, 20, 22, 25, 34]. Discriminative approaches are generally restricted to the viewpoints learned from training data. Though our dataset can be used for the training of discriminative methods, it will likely require retraining for each new applica-

tion. In contrast, our prior can be easily incorporated into generative approaches of pose estimation and tracking.

3. Pose-Conditioned Pose Prior

We observe that existing mocap datasets are not designed to explore pose-dependent joint angle limits. Consequently, we captured a new set of human motions performed by flexible people such as gymnasts and martial artists. Our capture protocol was designed to elicit a wide range of pairwise configurations of connected limbs in a kinematic tree (Fig. 4 (a)). We captured two types of movements. In the range of motion captures, participants were asked to keep their upper-arm fixed, fully flex and extend their lower-arms and then turn them inwards to outwards. This movement was repeated for a number of horizontal and vertical postures of the upper-arm. The same procedure was adopted for the legs. They were also asked to perform a number of stretching exercises (Fig. 1). From this data, we estimate a 17-point kinematic skeleton and learn joint angle limits.

We represent the human pose as a concatenation of 3D coordinates of P points $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T & \cdots & \mathbf{X}_P^T \end{bmatrix}^T \in \mathbb{R}^{3P\times 1}$. Let $\delta(.)$ be an operator that returns the relative coordinates of a joint with respect to its parent in the kinematic skeleton. We extend δ for vectors and matrices of points. The goal is to find a function

$$isvalid(\delta\mathbf{X}) : \mathbb{R}^{3\times N} \rightarrow \{0,1\}^N,$$

where N denotes the number of bones, and value 1 is returned if the corresponding bone is in a valid pose and 0 otherwise. Given a kinematic skeleton we first find a local coordinate system for each bone as we discuss next.

3.1. Global to Local Coordinate Conversion

In order to estimate joint-angle limits, we need to first find the local coordinate systems for all the joints. We can uniquely find a coordinate axis in 3D with respect to two non-parallel vectors $\mathbf{u}$ and $\mathbf{v}$. The three coordinate axes can be found using Gram-Schmidt on $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{u}\times\mathbf{v}$. We propose a conversion from $\delta\mathbf{X}$ to local coordinates $\tilde{\mathbf{X}}$ in Algorithm 1. For upper-arms, upper-legs and the head, $\mathbf{u}$ and $\mathbf{v}$ are defined with the help of the torso "bones" (spine, left/right hip, left/right shoulder) (lines 3-8). The selection of the coordinate system for every other bone, $\mathbf{b}$, is arbitrary and is defined with the help of an arbitrary vector, $\mathbf{a}$, and the parent bone, $\mathtt{pa}(\mathbf{b})$, of $\mathbf{b}$ (lines 10-11). $\mathbf{R_u}$ is the estimated rotation of this parent bone. Varying the values of the input vector, $\mathbf{a}$, can generate different coordinate systems and by keeping its value fixed we ensure consistency of the local coordinate system. Finally the local coordinate axes are found using Gram-Schmidt (line 12) and the local coordinates $\tilde{\mathbf{b}}$ are computed (line 13).

Algorithm 1 Global to Local Coordinate Conversion

1: Input $\delta\mathbf{X}$ and a constant arbitrary 3D vector $\mathbf{a}$.
2: **for** $\mathbf{b} \in \delta\mathbf{X}$
3: **if** ($\mathbf{b}$ is an upper-arm or head)
4: $\mathbf{u}$ = Left-shldr − Right-shldr;
5: $\mathbf{v}$ = back-bone;
6: **else if** ($\mathbf{b}$ is an upper-leg)
7: $\mathbf{u}$ = Left-hip − Right-hip;
8: $\mathbf{v}$ = back-bone;
9: **else**
10: $\mathbf{u} = \mathtt{pa}(\mathbf{b})$;
11: $\mathbf{v} = \mathbf{R_u a} \times \mathbf{u}$;
12: $\mathbf{R_b}$ = GramSchmidt$(\mathbf{u}, \mathbf{v}, \mathbf{u}\times\mathbf{v})$;
13: $\tilde{\mathbf{b}} = \mathbf{R_b^T b}$;
14: Return $\tilde{\mathbf{X}} = \{\tilde{\mathbf{b}}\}$;

3.2. Learning Joint-Angle Limits

We convert the local coordinates of the upper-arms, upper-legs and the head into spherical coordinates. Using our dataset, we then define a binary occupancy matrix for these bones in discretized azimuthal and polar angles, θ and ϕ respectively. A bone is considered to be in a valid position if its azimuthal and radial angles give a value 1 in the corresponding occupancy matrix (Fig. 3(a)).

The validity of every other bone $\mathbf{b}$ is decided conditioned on the position of its parent with a given θ and ϕ. Under this conditioning the bone can only lie on a hemisphere or even a smaller part of it. To exploit this we propose two types of constraints to check the validity of $\mathbf{b}$. First we find a half-space, $\mathbf{b}^T\mathbf{n} + d < 0$, defined by a separating plane with the normal vector $\mathbf{n}$ and the distance to origin d. Second we project all the instances of $\mathbf{b}$ in the dataset to the plane and find a bounding box enclosing these projections. A bone is considered to be valid if it lies in the half-space and its projection is inside the bounding-box (Fig. 3(b)). The separating plane is estimated by the following optimization,

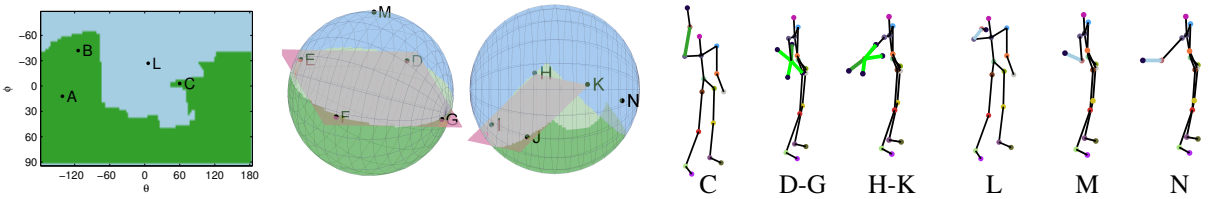
$$\min_{\mathbf{n},d} d^2 \quad \text{subject to} \quad \mathbf{A}^T\mathbf{n} < -d\mathbf{1}, \tag{1}$$

where $\mathbf{A}$ is a column-vise concatenation of all the instances of $\mathbf{b}$ in the dataset.

Figure 3 shows a visualization of our learned joint-angle limits. It shows that the joint angle limits for the wrist are different for two different positions of the elbow.

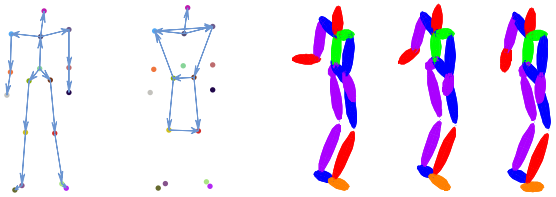
3.3. Augmenting 3D Pose Sparse Representation

To represent 3D pose, a sparse representation is proposed in [23], which uses a linear combination of basis poses



(a) elbow distribution (b) cond-wrist distribution for A & B (c) valid samples (points C to K) (d) invalid samples (L to N)

Figure 3. **Pose-dependent joint-angle limit.** (a) Occupancy matrix for right elbow in azimuthal and polar angles: green/sky-blue areas represent valid/invalid poses as observed in our capture data. (b) Given the elbow locations at A and B, the wrist can only lie on the green regions of the spheres. These valid wrist positions project to a box on the plane separating valid and invalid poses. The plots show that the valid poses of the wrist depend on the position of the elbow. (c) and (d) illustrate the valid (in green) and invalid (in sky-blue) elbow and wrist positions for the corresponding selected points in plots (a) and (b).



(a) $\delta\mathbf{X}$ (b) ex-torso (c) Effect of prior

Figure 4. Representation and ambiguity. (a) The δ operator computes relative coordinates by considering the parent as the origin. (b) The Bayesian network for the extended-torso exploits the relatively rigid locations of the joints within the torso and the correlation of left and right knee. (c) The over-complete dictionary representation allows invalid poses. Left to right: i) A 3D pose, where the right lower-arm violates the joint-angle limits is shown. ii) The over-complete dictionary represents this invalid 3D pose with a small number of basis poses (20 in comparison with the full dimensionality of 51). iii) Applying our joint-angle-limit prior makes the invalid pose valid.

$\mathbf{B}_1, \mathbf{B}_2, \cdots, \mathbf{B}_K$, plus the mean pose $\boldsymbol{\mu}$,

$$\hat{\mathbf{X}} = \boldsymbol{\mu} + \sum_{i=1}^{K} \omega_i \mathbf{B}_i = \boldsymbol{\mu} + \mathbf{B}^*\boldsymbol{\omega}, \qquad \{\mathbf{B}_i\}_{i \in \mathcal{I}_{B^*}} \in \mathbf{B}^* \subset \mathcal{B}, \tag{2}$$

where $\boldsymbol{\omega}$ is a vector of pose coefficients, ω_i, the matrix $\mathbf{B}^*$ is a column-wise concatenation of basis poses $\mathbf{B}_i$ selected with column indices $\mathcal{I}_{B^*}$ from an over-complete dictionary $\mathcal{B}$. $\mathcal{B}$ is computed by concatenating the bases of many actions and each basis is learned using Principal Component Analysis (PCA) on an action class. $\hat{\mathbf{X}}$ denotes the approximate 3D pose aligned with the basis poses and is related to the estimated pose, $\mathbf{X}$, by the camera rotation $\mathbf{R}$ as, $\mathbf{X} \approx (\mathbf{I}_{P\times P} \otimes \mathbf{R})\,\hat{\mathbf{X}}$. This sparse representation provides better generalization than PCA [23].

We observe that despite good generalization, the sparse representation also allows invalid poses. It is very easy to stay in the space spanned by the basis vectors, yet move outside the space of valid poses. Figure 4(c) shows that a small number of basis poses can reconstruct an invalid 3D pose, whereas our joint-angle-limit prior prevents the invalid configuration. We estimate this pose by solving the following optimization problem

$$\min_{\boldsymbol{\omega}} \|\mathbf{X} - (\mathbf{I} \otimes \mathbf{R})\,(\mathbf{B}^*\boldsymbol{\omega} + \boldsymbol{\mu})\,\|_2^2 + C_p, \tag{3}$$

where $\|.\|_2$ denotes the L^2 norm and where $C_p = 0$, if all the bones in $\delta\hat{\mathbf{X}}$ are valid according to the function $isvalid(.)$ and inf otherwise. Defining C_p this way is equivalent to adding nonlinear inequality constraints using the $isvalid(.)$ function.

4. 3D Pose Estimation

4.1. Preliminaries

Recall that human pose is represented as a concatenation of 3D coordinates of P points $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T & \cdots & \mathbf{X}_P^T \end{bmatrix}^T \in \mathbb{R}^{3P\times 1}$. Under a scaled orthographic camera model, the 2D coordinates of the points in the image are given by

$$\mathbf{x} = s\,(\mathbf{I}_{P\times P} \otimes \mathbf{R}_{1:2})\,\mathbf{X} + \mathbf{t} \otimes \mathbf{1}_{P\times 1}, \tag{4}$$

where $\mathbf{x} \in \mathbb{R}^{2P\times 1}$ and $s, \mathbf{R}$, and $\mathbf{t}$ denote the camera scale, rotation and translation parameters, $\otimes$ denotes the Kronecker product and the subscript $1:2$ gives the first two rows of the matrix. We can make $\mathbf{t} = \mathbf{0}$, under the assumption that the 3D centroid gets mapped to the 2D centroid and these are the origins of the world and the camera coordinate systems. Once the 3D pose is known, the actual value of $\mathbf{t}$ can be estimated using Equation (4).

Ramakrishna et al. [23] exploit the sparse representation in Equation (2) to find the unknown 3D pose $\mathbf{X}$. They minimize the following reprojection error to find $\boldsymbol{\omega}, I_{B^*}, s$, and $\mathbf{R}$ using a greedy Orthogonal Matching Pursuit (OMP) algorithm subject to an anthropometric regularization,

$$C_r(\boldsymbol{\omega}, \mathcal{I}_{B^*}, s, \mathbf{R}) = \|\mathbf{x} - s\,(\mathbf{I} \otimes \mathbf{R}_{1:2})\,(\mathbf{B}^*\boldsymbol{\omega} + \boldsymbol{\mu})\,\|_2^2. \tag{5}$$

Once $\boldsymbol{\omega}, \mathcal{I}_{B^*}$, and $\mathbf{R}$ are known, the pose is estimated as

$$\mathbf{X} = (\mathbf{I} \otimes \mathbf{R}) \left(\mathbf{B}^* \boldsymbol{\omega} + \boldsymbol{\mu}\right). \tag{6}$$

4.2. The Objective Function

Our method for 3D pose estimation given 2D joint locations exploits the proposed pose prior and the fact that bone-lengths follow known proportions. To learn the over-completely dictionary we choose the same CMU mocap sequences as were selected by Ramakrishna et al. [23] and add two further action classes "*kicks*" and "*pantomine*." To focus on pose and not body proportions we take the approach of Fan et al. [9] and normalize all training bodies to have the same mean bone length and all bodies to have the same proportions, giving every training subject the same bone lengths. We align the poses using Procrustes alignment of the extended-torso, defined below. We learn the PCA basis on each action class and concatenate the bases to get the over-complete dictionary. We also learn the PCA basis and the covariance matrix for the extended-torso, which we use for its pose estimation in the next section.

We estimate the 3D pose by minimizing,

$$\min_{\boldsymbol{\omega}, s, \mathbf{R}} C_r + C_p + \beta C_l, \tag{7}$$

where β is a normalization constant and the cost C_l penalizes the difference between the squares of the estimated i^{th} bone length $\|\delta(\hat{\mathbf{X}}_i)\|_2$ and the normalized mean bone length l_i, $C_l = \sum_{i=1}^{N} \left| \|\delta(\hat{\mathbf{X}}_i)\|_2^2 - l_i^2 \right|$, where $|.|$ denotes the absolute value and $\hat{\mathbf{X}}$ is estimated using Equation (2). We use an axis-angle representation to parameterize $\mathbf{R}$. We do not optimize for the basis vectors but estimate them separately as discussed Section 4.4.

An important consideration in minimizing the cost, given in Equation (7) as well as the objective function in previous methods [9, 22], is the sensitivity to initialization. In particular a good guess of the camera rotation matrix $\mathbf{R}$ is required to estimate the correct 3D pose. To solve this problem we notice that an extended-torso, consisting of the torso, head and upper-legs exhibits less diversity of poses than the full body and its pose estimation can give a more accurate estimate of the camera matrix.

4.3. Pose Estimation for Extended-Torso

To estimate the 3D pose for the extended-torso, we minimize a cost similar to Equation (7), but instead of the full-body, $\mathbf{X}$, we only consider points in the extended torso $\mathbf{X}'$. We learn a PCA basis $\mathbf{B}'$ for the extended torso with mean $\boldsymbol{\mu}'$. Hence a basis-aligned pose is given by, $\hat{\mathbf{X}}' = \mathbf{B}'\boldsymbol{\omega}' + \boldsymbol{\mu}'$.

Even the PCA-based modelling of the extended torso is not enough to constrain its 3D pose estimation from 2D. We model a prior on $\hat{\mathbf{X}}'$ by exploiting the inter-dependencies between points in the form of a Bayesian network (Fig. 4(b)). This network exploits the fact that the human torso is almost rigid and often left and right knees move in correlation. Hence, the probability of a pose is given by

$$p\left(\delta\hat{\mathbf{X}}'\right) = \prod_i p\left(\delta\hat{\mathbf{X}}'_i | \delta\hat{\mathbf{X}}'_{\mathcal{I}}\right), \tag{8}$$

where $\delta\hat{\mathbf{X}}'_{\mathcal{I}}$ denotes a vector obtained by concatenating the 3D coordinates of the points in the conditioning set defined by the Bayesian network. Under the assumption that the pair $\left(\delta\hat{\mathbf{X}}'_i, \delta\hat{\mathbf{X}}'_{\mathcal{I}}\right)$ is Gaussian distributed, we show in the Appendix that the prior on pose can be written as a linear constraint, $\mathbf{A}_p\boldsymbol{\omega}' = \mathbf{0}$, where $\mathbf{A}_p$ is computed using the basis $\mathbf{B}'$ and the covariance matrix of $\delta\hat{\mathbf{X}}'$. Hence, the prior term for the extended torso becomes, $C'_p = \|\mathbf{A}_p\boldsymbol{\omega}'\|_2^2$. We estimate the pose for the extended torso by minimizing the following objective analogous to Equation (7),

$$\min_{\boldsymbol{\omega}', s, \mathbf{R}} C'_r + \alpha C'_p + \beta C'_l. \tag{9}$$

We initialize the optimization by finding $\mathbf{R}$ and s using Procrustes alignment between the 2D joint locations $\mathbf{x}'$ and $\boldsymbol{\mu}'$. We find the solution using Quasi-Newton optimization. The estimated $\boldsymbol{\omega}'$, s, and $\mathbf{R}$ are used for the basis estimation for the full body in the next stage.

4.4. The Basis Estimation

Algorithm 2 Orthogonal Matching Pursuit (OMP)

$\mathbf{r}_{p0} = \mathbf{x} - s\left(\mathbf{I} \otimes \mathbf{R}_{1:2}\right)\boldsymbol{\mu}$;
$\mathbf{r}_{d0} = \delta\mathbf{Z}(\mathcal{I}_d) - s\left(\mathbf{I} \otimes \mathbf{R}_3\right)\delta\boldsymbol{\mu}(\mathcal{I}_d)$;
while $t < K$ **do**
  $i_{max} = \arg\max_i \big(\langle \mathbf{r}_{pt}, s\left(\mathbf{I} \otimes \mathbf{R}_{1:2}\right)\mathbf{B}_i \rangle + \langle \mathbf{r}_{dt}, s\left(\mathbf{I} \otimes \mathbf{R}_3\right)\delta\mathbf{B}_i(\mathcal{I}_d) \rangle \big)$;
  $\mathbf{B}^* = [\mathbf{B}^* \mathbf{B}_{i_{max}}]$;
  $\boldsymbol{\omega}^* = \arg\min_{\boldsymbol{\omega}} \big(\|\mathbf{x} - s\left(\mathbf{I} \otimes \mathbf{R}_{1:2}\right)\left(\mathbf{B}^*\boldsymbol{\omega} + \boldsymbol{\mu}\right)\|_2^2 + \|\delta\mathbf{Z}(\mathcal{I}_d) - s\left(\mathbf{I} \otimes \mathbf{R}_3\right)\left(\delta\mathbf{B}^*(\mathcal{I}_d)\boldsymbol{\omega} + \delta\boldsymbol{\mu}(\mathcal{I}_d)\right)\|_2^2 \big)$;
  $\mathbf{R} = \arg\min_{\mathbf{R}} \|\mathbf{x} - s\left(\mathbf{I} \otimes \mathbf{R}_{1:2}\right)\left(\mathbf{B}^*\boldsymbol{\omega}^* + \boldsymbol{\mu}\right)\|_2^2$;
  if $!isvalid\left(\delta\left(\mathbf{B}^*\boldsymbol{\omega} + \boldsymbol{\mu}\right)\right)$
    remove $\mathbf{B}_{i_{max}}$ and **go to** step 4
  $\mathbf{r}_{pt} = \mathbf{x} - s\left(\mathbf{I} \otimes \mathbf{R}_{1:2}\right)\left(\mathbf{B}^*\boldsymbol{\omega}^* + \boldsymbol{\mu}\right)$;
  $\mathbf{r}_{dt} = \delta\mathbf{Z}(\mathcal{I}_d) - s\left(\mathbf{I} \otimes \mathbf{R}_3\right)\left(\delta\mathbf{B}^*(\mathcal{I}_d)\boldsymbol{\omega}^* + \delta\boldsymbol{\mu}(\mathcal{I}_d)\right)$;
Return $\{\mathbf{R}, \mathbf{B}^*\}$;

In this step we estimate the basis $\mathbf{B}^*$ using an OMP algorithm similar to Ramakrishna et al. [23]. The difference is that here we already know the depth of a few of the bones by exploiting the joint-angle limit constraints. Additionally, we do not impose a hard constraint that the bone lengths have to sum to a predefined number.

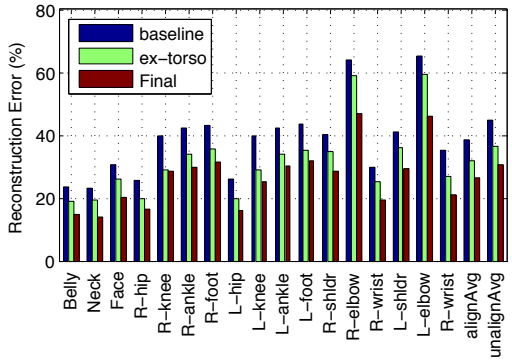


Figure 5. Impact of the extended-torso initialization and the proposed pose prior: Reconstruction error is average Euclidean distance per joint between the estimated and the ground-truth 3D pose and is measured as a fraction of the back-bone length. Error decreases monotonically with the addition of each module.

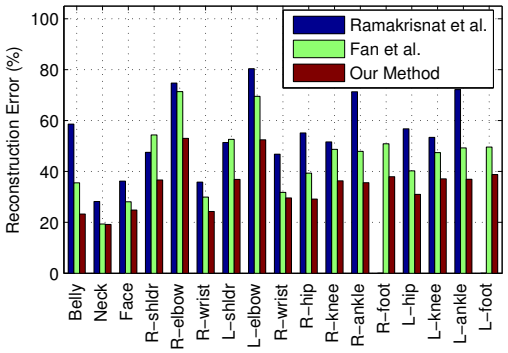


Figure 6. The proposed method gives consistently smaller reconstruction error in comparison with the other two methods.

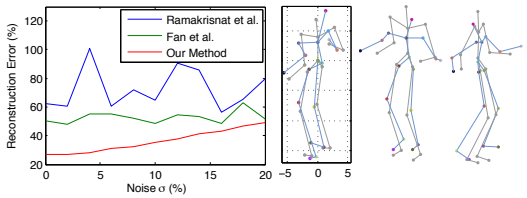


Figure 7. The proposed method is robust to a fairly large range of noise in comparison with the previous methods. Noise σ is proportional to the back-bone length. A sample input frame at $\sigma = 20\%$ and our estimated 3D pose with error=50% is also shown in two views (gray: ground-truth, colored: estimated).

Let $\mathbf{Z}$ denote the vector of unknown depths of all the points in 3D pose. Given the mean bone lengths l_i and the estimated orthographic scale s, we estimate the absolute relative depths $|\delta\mathbf{Z}|$ using Taylor's method [29]. Since natural human poses are not completely arbitrary, the unknown signs of the relative depths can be estimated for some of the bones by exploiting joint-angle limits. We generate all signs of the bones in an arm or leg and test whether they correspond to a valid pose using the function $isvalid(\delta\mathbf{X})$. The sign of a bone is taken to be positive if, according to our prior, a negative sign is not possible in any of the combinations for the corresponding arm or leg. If not positive, we do the same test in the other direction to see if the sign can be negative. If neither is possible, we must rely on the overcomplete basis. The indices of the depths estimated this way are denoted as $\mathcal{I}_d$.

Given the 2D joint locations, $\mathbf{x}$, the relative depths estimated above, $\delta\mathbf{Z}(\mathcal{I}_d)$, the current estimate of s and $\mathbf{R}$, OMP, given in Algorithm 2, proceeds in a greedy fashion. The algorithm starts with a current estimate of 3D pose as $\boldsymbol{\mu}$ and computes the initial residual for the 2D projection and known relative depths (line 1,2). At each iteration a basis vector from $\mathcal{B}$ is chosen and added to $\mathbf{B}^*$ that is most aligned with the residual under the current estimate of rotation (line 4,5). Then given $\mathbf{B}^*$, the pose coefficients $\boldsymbol{\omega}^*$ and camera rotations $\mathbf{R}$ are re-estimated (line 6,7). We remove the basis vector if it makes the resulting pose invalid and consider the basis vector with the next highest dot product (line 8,9). The residual is updated using $\mathbf{B}^*$, $\boldsymbol{\omega}^*$, and the new estimate of $\mathbf{R}$ (line 10,11). The algorithm terminates when $\mathbf{B}^*$ has reached a predefined size.

Finally, the estimated $\mathbf{B}^*$, $\boldsymbol{\omega}^*$, and $\mathbf{R}$ are used to initialize the optimization in Equation (7).

5. Experiments

We compare the pose-prior learned from our dataset with the same prior learned from the CMU dataset. We classify all the poses in our dataset as valid or invalid using the prior learned from CMU. We find that out of a 110 minutes of data about 12% is not explained by the CMU-based prior. This suggests that the CMU dataset does not cover the full range of human motions. A similar experiment shows that out of 9.5 hours of CMU data about 8% is not explained by our prior. A closer investigation reveals that CMU contains many mislabeled markers. This inflates the space of valid CMU poses to include invalid ones. Removing the invalid poses would likely increase the percentage of our poses that are not explained and would decrease the amount of CMU data unexplained by our prior.

We quantitatively evaluate our method using all CMU mocap sequences of four actors (103, 111, 124, and 125) for a total of 69 sequences. We create two sets of synthetic images, called testset1 and testset2, by randomly selecting 3000 and 10000 frames from these sequences and projecting them using random camera viewpoints. We report reconstruction error per joint as the average Euclidean distance between the estimated and the ground-truth pose. Like previous methods [9, 23] we Procrustes align the estimated 3D pose with the ground-truth to compute the error. To fix arbitrary scale, we divide the ground by back-bone

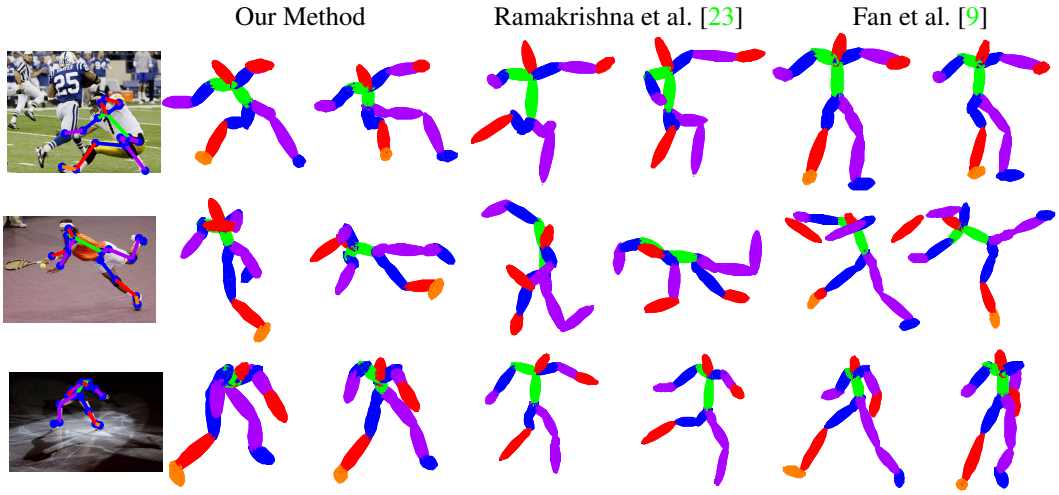


Figure 8. Real results with manual annotation. We demonstrate substantial improvement over the previous methods. The proposed method gives an anthropometrically valid interpretation of 2D joint locations whereas the previous methods often give invalid 3D poses.

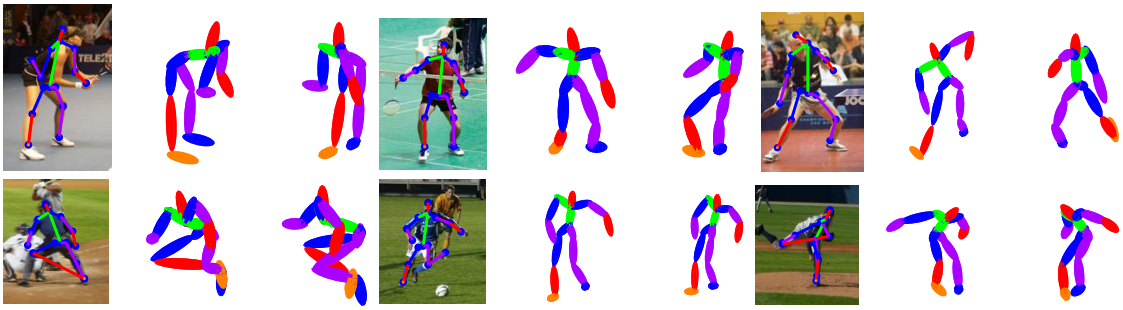


Figure 9. Real results with automatic part-based detections [16] on the Leeds sports pose dataset on a few frames. Despite the outliers in detections, our method gives valid 3D pose interpretations. Please note that feet were not detected in the images but with the help of our pose prior their 3D location is estimated.

length and Procrustes align this with the estimated pose. We also evaluate the camera matrix estimation.

We first evaluate the impact of the extended-torso initialization and joint-angle prior in the proposed method on testset1. We start with a baseline consisting of just the projected matching pursuit algorithm and test its accuracy. We initialize this by finding $\mathbf{R}$ and s by Procrustes alignment between $\mathbf{x}$ and $\boldsymbol{\mu}$. Then we include the initialization using pose estimation for the extended-torso and the final joint optimization to enforce length constraints. Finally, we include the depth estimation using joint-angle limits and the proposed pose prior in the joint optimization. In Fig. 5 we report the mean reconstruction errors per joint for this experiment. The results show a monotonic decrease in error with the addition of each of these modules. We also report the overall mean reconstruction error with and without Procrustes alignment. For the later case we multiply the camera rotation with the 3D pose and adopt a canonical camera convention. Observing that both the errors are roughly equal we conclude that the estimated camera matrices are correct.

Next we compare the accuracy of our method against the previous methods on testset2. The source code for the previous methods were kindly provided by the authors. Note that the method by Fan et al. is customized for a few classes of actions, including walking, running, jumping, boxing, and climbing and its accuracy is expected to degrade on other types of actions. Figure 6 shows that the proposed method outperforms the other two methods. In Fig. 7 we test the sensitivity of our algorithm against Gaussian noise and compare it against the methods by Ramakrishna et al. [23] and Fan et al. [9]. We add noise proportional to the backbone length in 3D, project the noisy points using random camera matrices and report our pose estimation accuracy. The results demonstrate that the proposed method is significantly more robust than the previous methods. Our experiments show that the proposed method gives a small reprojection error and an anthropometrically valid 3D pose interpretation, whereas the previous methods often estimate an invalid 3D pose. A further investigation reveals that the reconstruction error in canonical camera convention for the previous methods is significantly worse than the one with Procrustes alignment (55% and 44% vs. 143% and 145%

respectively), whereas for our method the errors are not significantly different (34% vs. 45%). This implies that an important reason for the failure of previous methods is the incorrect estimation of the camera matrix. This highlights the contribution of extended-torso initialization.

It is important to mention the inherent ambiguities in 3D pose estimation (see Fig. 2), which imply that given 2D point locations, a correct pose estimation can never be insured and only a probable 3D pose can be estimated. Results show that the proposed method satisfies this criterion.

Figure 8 shows results on real images with manual annotations of joints and compares them with the previous methods by showing the 3D pose in two arbitrary views. Again the results show that our method gives a valid 3D pose whereas the previous methods often do not. Figure 9 shows results with automatic part-based detections [16] on the Leeds sports pose dataset on a few frames. Results show that despite significant noise in detection, the proposed method is able to recover a valid 3D pose. For more results please see supplementary material [1].

6. Conclusion

We propose pose-conditioned joint angle limits and formulate a prior for human pose. We believe that this is the first general prior to consider pose dependency of joint limits. We demonstrate that this prior restricts invalid poses in 2D-to-3D human pose reconstruction. Additionally we provide a new algorithm for estimating 3D pose that exploits our prior. Our method significantly outperforms the current state of the art methods both quantitatively and qualitatively.

Our prior and the optimization framework can be applied to many problems in human pose estimation beyond the application described here. In future we will consider dependencies of siblings in the kinematic tree on joint-angle limits. We are also working on temporal models of 2D-to-3D pose estimation that further reduce ambiguities. Future work should also consider the temporal dependency of joint limits since, during motion, the body can reach states that may not be possible statically.

7. Appendix

We model the pose prior on the extended-torso as the following Bayesian network,

$$p\left(\hat{\mathbf{X}}\right) = \prod_i p\left(\delta\hat{\mathbf{X}}'_i | \delta\hat{\mathbf{X}}'_{\mathcal{I}}\right), \tag{10}$$

where $\mathcal{I}$ denotes the indices of the joints in the conditioning set defined by the Bayesian network shown in Fig. 4(b) and $\delta\hat{\mathbf{X}}'_{\mathcal{I}}$ is a vector obtained by concatenating their 3D coordinates. We consider the combined Gaussian distribution of a joint i and its conditioning set as,

$$\begin{pmatrix} \delta\hat{\mathbf{X}}'_i \\ \delta\hat{\mathbf{X}}'_{\mathcal{I}} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \delta\boldsymbol{\mu}'_i \\ \delta\boldsymbol{\mu}'_{\mathcal{I}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}'_{ii} & \boldsymbol{\Sigma}'_{i\mathcal{I}} \\ \boldsymbol{\Sigma}'_{\mathcal{I}i} & \boldsymbol{\Sigma}'_{\mathcal{I}\mathcal{I}} \end{pmatrix}\right), \tag{11}$$

where the relative pose $\delta\hat{\mathbf{X}}'$ satisfies,

$$\delta\hat{\mathbf{X}}' = \delta\mathbf{B}'\boldsymbol{\omega}' + \delta\boldsymbol{\mu}'. \tag{12}$$

Given Equation (11) the conditional distribution can be written as, $\left(\delta\hat{\mathbf{X}}'_i | \delta\hat{\mathbf{X}}'_{\mathcal{I}} = \mathbf{a}\right) \sim \mathcal{N}\left(\delta\overline{\boldsymbol{\mu}}'_i, \overline{\boldsymbol{\Sigma}}'_{ii}\right)$, where

$$\begin{aligned} \delta\overline{\boldsymbol{\mu}}'_i &= \delta\boldsymbol{\mu}'_i + \boldsymbol{\Sigma}'_{i\mathcal{I}}\boldsymbol{\Sigma}'^{-1}_{\mathcal{I}\mathcal{I}}(\mathbf{a} - \delta\boldsymbol{\mu}'_{\mathcal{I}}), \\ \overline{\boldsymbol{\Sigma}}'_{ii} &= \boldsymbol{\Sigma}'_{ii} - \boldsymbol{\Sigma}'_{i\mathcal{I}}\boldsymbol{\Sigma}'^{-1}_{\mathcal{I}\mathcal{I}}\boldsymbol{\Sigma}'_{\mathcal{I}i}. \end{aligned} \tag{13}$$

The above pose prior can be combined with Equation (12) by noticing that $(\mathbf{a} - \delta\boldsymbol{\mu}'_{\mathcal{I}}) = \delta\mathbf{B}'_{\mathcal{I}}\boldsymbol{\omega}'$, where $\delta\mathbf{B}'_{\mathcal{I}}$ consists of the rows from $\delta\mathbf{B}'$ corresponding to the points $\mathcal{I}$. Using this relation the complete vector $\delta\overline{\boldsymbol{\mu}}'$ can be estimated as,

$$\delta\overline{\boldsymbol{\mu}}' - \delta\boldsymbol{\mu}' = \mathbf{G}\boldsymbol{\omega}', \tag{14}$$

where $\mathbf{G}$ is formed by stacking the matrices $\boldsymbol{\Sigma}'_{i\mathcal{I}}\boldsymbol{\Sigma}'^{-1}_{\mathcal{I}\mathcal{I}}\delta\mathbf{B}'_{\mathcal{I}}$ for all i. Equation (14) provides the mean 3D pose under the Gaussian network prior. The covariance of pose $\overline{\boldsymbol{\Sigma}}'$ is formed by stacking all conditional covariances $\overline{\boldsymbol{\Sigma}}'_{ii}$ from Equation (13). This prior on 3D pose is used to formulate a prior on $\boldsymbol{\omega}' \sim \mathcal{N}\left(\overline{\boldsymbol{\mu}}_{\omega'}, \overline{\boldsymbol{\Sigma}}_{\omega'}\right)$ using Equation (12) as,

$$\overline{\boldsymbol{\mu}}_{\omega'} = \delta\mathbf{B}'^{\dagger}\mathbf{G}\boldsymbol{\omega}', \quad \text{and} \quad \overline{\boldsymbol{\Sigma}}_{\omega'} = \delta\mathbf{B}'^{\dagger}\overline{\boldsymbol{\Sigma}}'\delta\mathbf{B}'^{\dagger T}, \tag{15}$$

where the superscript $\dagger$ denotes MoorePenrose pseudoinverse. This prior can be used to formulate a MAP estimate of $\boldsymbol{\omega}'$. The likelihood equation for $\boldsymbol{\omega}'$ is the following,

$$\mathbf{x}' = s\left(\mathbf{I} \otimes \mathbf{R}_{1:2}\right)\left(\mathbf{B}'\boldsymbol{\omega}' + \boldsymbol{\mu}'\right). \tag{16}$$

The above becomes linear if the camera matrix and orthographic scale-factor are known and can be written as a matrix multiplication, $\mathbf{A}\boldsymbol{\omega}' = \mathbf{b}$. Therefore the likelihood distribution can be written as $\mathbf{b}|\boldsymbol{\omega}' \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\omega}', \alpha'\mathbf{I}\right)$, where α' is the variance of the noise. Using this the MAP estimate of $\boldsymbol{\omega}$ can be found by minimizing the sum of Mahalanobis distances of both the prior and likelihood distributions,

$$c(\boldsymbol{\omega}') = \frac{1}{\alpha}\|\mathbf{A}\boldsymbol{\omega}' - \mathbf{b}\|^2 + \left(\boldsymbol{\omega}' - \overline{\boldsymbol{\mu}}_{\omega'}\right)^T \overline{\boldsymbol{\Sigma}}^{-1}_{\omega'}\left(\boldsymbol{\omega}' - \overline{\boldsymbol{\mu}}_{\omega'}\right).$$

By taking the partial derivatives of c with respect to $\boldsymbol{\omega}'$, a linear system of equations can be made and the MAP estimate of $\boldsymbol{\omega}'$ can be found as the following,

$$\left(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{D}^T\overline{\boldsymbol{\Sigma}}^{-1}_{\omega'}\mathbf{D}\right)\boldsymbol{\omega}' = \mathbf{A}^T\mathbf{b}, \tag{17}$$

where $\mathbf{D} = \mathbf{I} - \delta\mathbf{B}'^{\dagger}\mathbf{G}$. Solving this linear system is equivalent to solving two sets of equations, $\mathbf{A}\boldsymbol{\omega}' = \mathbf{b}$, and

$$\sqrt{\alpha}\mathbf{A}_p\boldsymbol{\omega}' = \mathbf{0}, \tag{18}$$

where $\mathbf{A}_P$ is a Cholesky decomposition of $\mathbf{D}^T\overline{\boldsymbol{\Sigma}}^{-1}_{\omega'}\mathbf{D}$. We use Equation (18) to add a prior for the extended-torso.

8. Acknowledgements

We thank Andrea Keller, Sophie Lupas, Stephan Streuber, and Naureen Mahmood for joint-limit dataset preparation. We benefited from discussions with Jonathan Taylor, Peter Gehler, Gerard Pons-Moll, Varun Jampani, and Kashif Murtza. We also thank Varun Ramakrishna and Xiaochuan Fan for providing us their source code.

References

[1] `http://poseprior.is.tue.mpg.de/`.

[2] U. S. N. Aeronautics and S. Administration. *NASA-STD-3000: Man-systems integration standards*. Number v. 3 in NASA-STD. National Aeronautics and Space Administration, 1995.

[3] M. Andriluka, S. Roth, and B. Schiele. Monocular 3D pose estimation and tracking by detection. In *Computer Vision and Pettern Recognition*, pages 623–630, 2010.

[4] C. Barròn and I. Kakadiaris. Estimating anthropometry and pose from a single uncalibrated image. *Computer Vision and Image Understanding*, 81(3):269–284, March 2001.

[5] C. BenAbdelkader and Y. Yacoob. Statistical estimation of human anthropometry from a single uncalibrated image. In *Methods, Applications, and Challenges in Computer-assisted Criminal Investigations*, Studies in Computational Intelligence. Springer-Verlag, 2008.

[6] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *International Conference on Computer Vision*, pages 1365–1372, Sept. 2009.

[7] J. Chen, S. Nie, and Q. Ji. Data-free prior model for upper body pose estimation and tracking. *IEEE Trans. Image Proc.*, 22(12):4627–4639, Dec. 2013.

[8] M. G. Choi, K. Yang, T. Igarashi, J. Mitani, and J. Lee. Retrieval and visualization of human motion data via stick figures. *Computer Graphics Forum*, 31(7):2057–2065, 2012.

[9] X. Fan, K. Zheng, Y. Zhou, and S. Wang. Pose locality constrained representation for 3d human pose reconstruction. In *Computer Vision–ECCV 2014*, pages 174–188. Springer, 2014.

[10] K. Grochow, S. L. Martin, A. Hertzmann, and Z. Popović. Style-based inverse kinematics. *ACM Transactions on Graphics (TOG)*, 23(3):522–531, 2004.

[11] P. Guan, A. Weiss, A. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *Int. Conf. on Computer Vision, ICCV*, pages 1381–1388, Sept. 2009.

[12] H. Hatze. A three-dimensional multivariate model of passive human joint torques and articular boundaries. *Clinical Biomechanics*, 12(2):128–135, 1997.

[13] S. Hauberg, S. Sommer, and K. Pedersen. Gaussian-like spatial priors for articulated tracking. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, volume 6311 of *Lecture Notes in Computer Science*, pages 425–437. Springer Berlin Heidelberg, 2010.

[14] L. Herda, R. Urtasun, and P. Fua. Hierarchical implicit surface joint limits for human body tracking. *Computer Vision and Image Understanding*, 99(2):189–209, 2005.

[15] H. Jiang. 3D human pose reconstruction using millions of exemplars. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 1674–1677. IEEE, 2010.

[16] M. Kiefel and P. Gehler. Human pose estimation with fields of parts. In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 331–346. Springer International Publishing, Sept. 2014.

[17] T. Kodek and M. Munich. Identifying shoulder and elbow passive moments and muscle contributions. In *IEEE Int. Conf. on Intelligent Robots and Systems*, volume 2, pages 1391–1396, 2002.

[18] H. J. Lee and Z. Chen. Determination of 3D human body postures from a single view. *Computer Vision Graphics and Image Processing*, 30(2):148–168, 1985.

[19] J. Lin, T. Igarashi, J. Mitani, M. Liao, and Y. He. A sketching interface for sitting pose design in the virtual environment. *Visualization and Computer Graphics, IEEE Transactions on*, 18(11):1979–1991, 2012.

[20] G. Mori and J. Malik. Recovering 3D human body configurations using shape contexts. *Pattern Analysis and Machine Intelligence*, 28(7):1052–1062, 2006.

[21] V. Parameswaran and R. Chellappa. View independent human body pose estimation from a single perspective image. In *Computer Vision and Pettern Recognition*, pages 16–22, 2004.

[22] I. Radwan, A. Dhall, and R. Goecke. Monocular image 3D human pose estimation under self-occlusion. In *International Conference on Computer Vision*, pages 1888–1895, 2013.

[23] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3D human pose from 2D image landmarks. *European Conference on Computer Vision*, pages 573–586, 2012.

[24] J. M. Rehg, D. D. Morris, and T. Kanade. Ambiguities in visual tracking of articulated objects using two-and three-dimensional models. *The International Journal of Robotics Research*, 22(6):393–418, 2003.

[25] G. Rogez, J. Rihan, C. Orrite-Uruñuela, and P. H. Torr. Fast human pose detection using randomized hierarchical cascades of rejectors. *International Journal of Computer Vision*, 99(1):25–52, 2012.

[26] M. Schünke, E. Schulte, and U. Schumacher. *Prometheus: Allgemeine Anatomie und Bewegungssystem : LernAtlas der Anatomie*. Prometheus LernAtlas der Anatomie. Thieme, 2005.

[27] C. Sminchisescu and B. Triggs. Building roadmaps of local minima of visual models. In *European Conference on Computer Vision*, volume 1, pages 566–582, Copenhagen, 2002.

[28] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6):371–391, 2003.

[29] C. J. Taylor. Reconstruction of articulated objects from point correspondences in a single uncalibrated image. *Computer Vision and Image Understanding*, 80(10):349–363, October 2000.

[30] J. Valmadre and S. Lucey. Deterministic 3D hman pose estimation using rigid structure. In *European Conference on Computer Vision*, pages 467–480. Springer, 2010.

[31] X. K. Wei and J. Chai. Modeling 3D human poses from uncalibrated monocular images. In *International Conference on Computer Vision*, pages 1873–1880, 2009.

[32] X. K. Wei and J. Chai. Intuitive interactive human-character posing with millions of example poses. *Computer Graphics and Applications, IEEE*, 31(4):78–88, 2011.

[33] I. Yoo, J. Vanek, M. Nizovtseva, N. Adamo-Villani, and B. Benes. Sketching human character animations by composing sequences from large motion database. *The Visual Computer*, 30(2):213–227, 2014.

[34] T.-H. Yu, T.-K. Kim, and R. Cipolla. Unconstrained monocular 3D human pose estimation by action detection and cross-modality regression forest. In *Computer Vision and Pettern Recognition*, pages 3642–3649, 2013.