Generating Multiple Diverse Hypotheses for Human 3D Pose Consistent with 2D Joint Detections

Ehsan Jahangiri, Alan L. Yuille
Johns Hopkins University, Baltimore, USA
ejahang1@jhu.edu, alan.yuille@jhu.edu

Abstract

We propose a method to generate multiple diverse and valid human pose hypotheses in 3D all consistent with the 2D detection of joints in a monocular RGB image. We use a novel generative model uniform (unbiased) in the space of anatomically plausible 3D poses. Our model is compositional (produces a pose by combining parts) and since it is restricted only by anatomical constraints it can generalize to every plausible human 3D pose. Removing the model bias intrinsically helps to generate more diverse 3D pose hypotheses. We argue that generating multiple pose hypotheses is more reasonable than generating only a single 3D pose based on the 2D joint detection given the depth ambiguity and the uncertainty due to occlusion and imperfect 2D joint detection. We hope that the idea of generating multiple consistent pose hypotheses can give rise to a new line of future work that has not received much attention in the literature. We used the Human3.6M dataset for empirical evaluation.

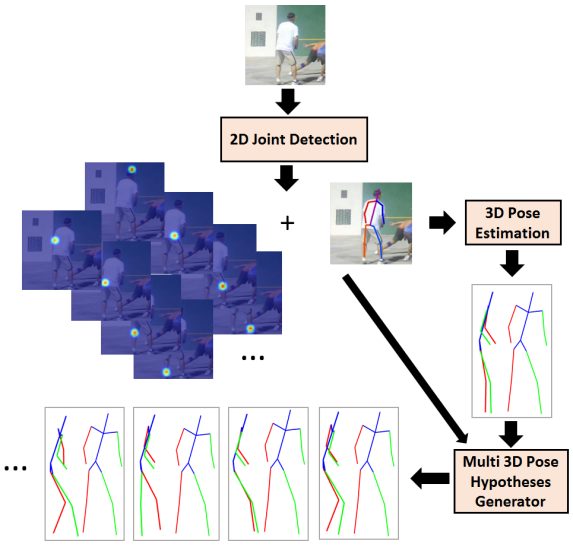


Figure 1. The input monocular image is first passed through a CNN-based 2D joint detector which outputs a set of heatmaps for soft localization of 2D joints. The 2D detections are then passed to a 2D-to-3D pose estimator to obtain an estimate of the 3D torso and the projection matrix. Using the estimated 3D torso, the projection matrix, and the output of the 2D detector we generate multiple diverse 3D pose hypotheses consistent with the output of 2D joint detector.

1. Introduction

Estimating the 3D pose configurations of complex articulated objects such as humans from monocular RGB images is a challenging problem. There are multiple factors contributing to the difficulty of this critical problem in computer vision: (1) multiple 3D poses can have similar 2D projections. This renders 3D human pose reconstruction from its projected 2D joints an ill-posed problem; (2) the human motion and pose space is highly nonlinear which makes pose modeling difficult; (3) detecting precise location of 2D joints is challenging due to the variation in pose and appearance, occlusion, and cluttered background. Also, minor errors in the detection of 2D joints can have a large effect on the reconstructed 3D pose. These factors favor a 3D pose estimation system that takes into account the uncertainties and suggests multiple possible 3D poses constrained only by reliable evidence. Often in the image, there exist much more detailed information about the 3D pose of a human than the 2D location of the joints (such as contextual information and difference in shading/texture due to depth disparity). Hence, most of the possible 3D poses consistent with the 2D joint locations can be rejected based on more detailed image information (*e.g.* in an analysis-by-synthesis framework or by investigating the image with some mid-level queries such as "Is the left hand in front of torso?") or by physical laws (*e.g.* gravity). We can also imagine scenarios where the image does not contain enough information to rule out or favor one 3D pose configuration over another especially in the presence of occlusion. In this paper, we focus on generating multiple plausible and diverse 3D pose hypotheses which while satisfying humans anatomical constraints are still consistent with the output of the 2D joint

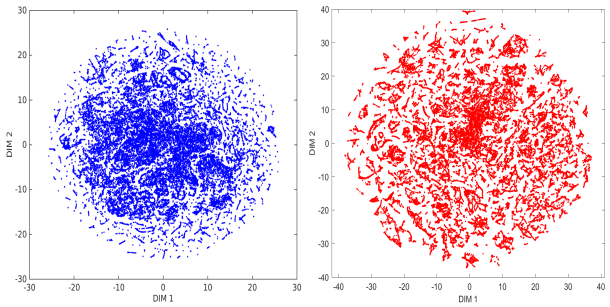
detector. Figure 1 illustrates an overview of our approach.

The space of valid human poses is a non-convex complicated space constrained by the anatomical and anthropomorphic limits. A bone never bends beyond certain angles with respect to its parent bone in the kinematic chain and its normalized length, with respect to other bones, cannot be much shorter/longer than standard values. This inspired Akhter and Black [1] to build a motion capture dataset composed of 3D poses of flexible subjects such as gymnasts and martial artists to study the joint angle limits. The statistics of 3D poses in this motion capture dataset is different from the previously existing motion capture datasets such as CMU [11], Human 3.6M [15], and HumanEva [28], because of their intention to explore the joint angle limits rather than performing and recognizing typical human actions. Figure 2 shows the t-SNE visualization [36] of poses from **A**khter&**B**lack motion **C**apture **D**ataset (**ABCD**) versus H36M in two dimensions. One can see that the "ABCD" dataset is more uniformly distributed compared to the H36M dataset. We randomly selected 4 poses from the dense and surrounding sparse areas in the H36M t-SNE map and have shown the corresponding images. One can see that all of the four samples selected from the dense areas correspond to standing poses whereas all of the four samples selected from sparse areas correspond to sitting poses.

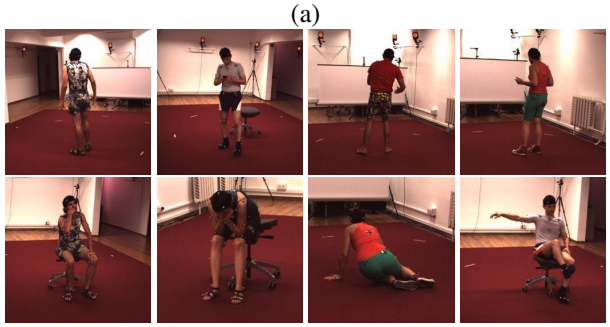
Training and testing a 3D model on a similarly biased dataset with excessive repetition of some poses will result in reduced performance on novel or rarely seen poses. As a simple demonstration, we learned a GMM 3D pose model [29] from a uniformly sampled set of Human 3.6M poses (all 15 actions) and evaluated the likelihood of 3D poses per action under this model. The average likelihood per action (up to a scaling factor) was: Directions 0.63, Discussion 0.74, Eating 0.56 , Greeting 0.63 , Phoning 0.28 , Posing 0.38 , Purchases 0.55 , Sitting 0.07 , Sitting Down 0.07 , Smoking 0.47 , Taking Photo 0.23 , Waiting 0.33 , Walking 0.64 , Walking Dog 0.29 , and Walk Together 0.25. According to the GMM model, the "Discussion" poses are on average almost 10 times more likely than "Sitting" poses which is due to the dataset and consequently the model bias. The EM algorithm used to learn the GMM model attempts to maximize the likelihood of all samples which will lead to a biased model if the training dataset is biased. Obviously, any solely data-driven model learned from a biased dataset that does not cover the full range of motion of human body can suffer from lack of generalization to novel or rarely seen yet anatomically plausible poses.

We propose a novel generative model on human 3D poses uniform in the space of physically valid poses (satisfying the constraints from [1]). Since our model is constrained only by the anatomical limits of human body it does not suffer from dataset bias which is intrinsically helpful to diversify pose hypotheses. Note that the pose-conditioned anatomical constraints calculated in [1] was originally used in a constrained optimization framework for single 3D pose estimation and turning those constraints into a generative model to produce uniform samples is not trivial. One of our main contributions is a pose-conditioned generative model which has not been done previously. We generate multiple anatomically-valid and diverse pose hypotheses consistent with the 2D joint detections to investigate the importance of having multiple pose hypotheses under depth and missing-joints (*e.g.* caused by occlusion) ambiguities. In the recent years, we have witnessed impressive progress in accurate 2D pose estimation of human in various pose and appearances which is made possible thanks to deep neural networks and lots of annotated 2D images. We take advantage of the recent advancement in human 2D pose estimation and seed our multi-hypotheses pose generator by an off-the-shelf 3D pose estimator. Namely, we use the "Stacked Hourglass" 2D joint detector [19] and the 2D-to-3D pose estimators of Akhter&Black [1] and Zhou et al. [42] to estimate the 3D torso and projection matrix. However, note that to our generic approach does not rely on any specific 2D/3D pose estimator and can easily adopt various 2D/3D pose estimators.

After briefly discussing some related works in subsection 1.1 we propose our approach in section 2. Our experimental results based on multiple 3D pose estimation baselines is given in section 3. We conclude in section 4.



(a)



(b)

Figure 2. (a): The t-SNE visualization of poses from the H36M (fist from left) and ABCD (second from left). (b): The images corresponding to the random selection of poses from the dense (top row in right) and sparse (bottom row in right) area of the H36M t-SNE map confirm the dataset bias toward standing poses compared to sitting poses.

1.1. Related Work

There are quite a few works in the human pose estimation literature that are directly or indirectly related to our work. Reviewing the entire literature is obviously beyond the scope of this paper. Several areas of research are related to our work such as 2D joiont detection, 3D pose estimation, and generative 3D pose modeling. Due to the advancements made by deep neural networks, the most recent works on 2D joint detection are based on convolutional neural networks (CNN) [35, 9, 34, 10, 40, 39, 38, 19, 6, 26] compared to the traditional hand-crafted feature based methods [27, 41, 12]. On the other hand, most of the 3D pose estimation methods use sparse coding based on an overcomplete dictionary of basis poses to represent a 3D pose and fit the 3D pose projection to the 2D joint detections [24, 37, 1, 42, 43]. Some works [8, 25, 26] try to train a deep network to directly predict 3D poses. However, purely discriminative approaches for 3D structure prediction (such as [8]) are usually very sensitive to data manipulation. On the other hand, it has been shown that the deep networks are very effective and more robust at detecting 2D templates (compared to 3D structures) such as human 2D body parts in images [19].

We use conditional sampling from our generative model to generate multiple consistent pose hypotheses. A number of previous works [7, 30, 2, 4, 5] have used sampling for human pose estimation. However, the sampling performed by these works are for purposes different from our goal to generate multiple diverse and valid pose hypotheses. For example, Amin et al. [2] use a mixture of pictorial structures and perform inference in two stages where the first stage reduces the search space for the second inference stage by generating samples for the 2D location of each part.

Some more closely related works include [33, 22, 16, 20, 23, 31, 17, 32]. Sminchisescu and Triggs [33] search for multiple local minima of their fitting cost function using a sampling mechanism based on forwards/backwards link flipping to generate pose candidates. Pons-Moll et al. [22] use inverse kinematics to sample the pose manifold restricted by the input video and IMU sensor cues in a particle filter framework. Lee and Cohen [16] use proposal maps to consolidate the evidence and generating 3D pose candidates during the MCMC search where they model the measurement uncertainty of 2D position of joints using a Gaussian distribution. Their MCMC approach suffers from high computational cost. Park and Ramanan [20] generate non-overlapping diverse pose hypotheses (only in 2D) from a part model. One interesting work is the "Posebit" by Pons-Moll et al. [23] that can retrieve pose candidates from a MoCap dataset of 3D poses given answers to some mid-level queries such as "Is the right hand in front of torso?" using decision trees. This approach is heavily dependent on the choice of MoCap dataset and cannot generalize to unseen poses. Simo-Serra1 et al. [31] model the 2D and 3D poses jointly in a Bayesian framework by integrating a generative model and discriminative 2D part detectors based on HOGs. Lehrmann et al. [17] learn a generative model from the H36M MoCap dataset whose graph structure (not a Kinematic chain) is learned using the Chow-Liu algorithm. Simo-Serra et al. [32] propagate the error in the estimation of 2D joint locations (modeled using Gaussian distributions) into the weights of dictionary elements in a sparse coding framework; then by sampling the weights, some 3D pose samples are generated and sorted based on the SVM score on joint distance features. However, their approach does not guarantee that the joint angle constraints are satisfied and do not address the depth ambiguity. We impose "pose-conditioned" joint angle and bone length constrains to ensure pose validity of samples from our generative model which has not been done before. In addition, our unbiased generative model restricted only by anatomical constrains helps in generating more diverse 3D pose hypotheses.

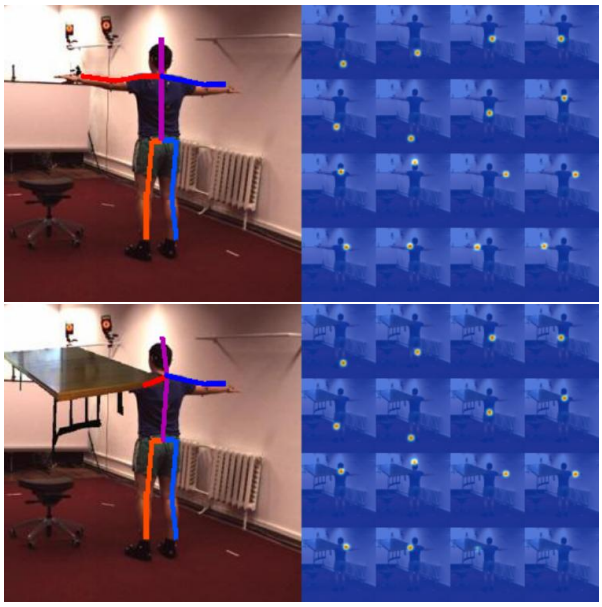


Figure 3. "Stacked Hourglass" 2D joint detector [19] in the absence and presence of occlusion. On the right-hand-side of each image are the corresponding heatmaps for joints.

2. The Proposed Method

Since our approach is closely related to the joint-angle constraints used in [1], we find it helpful for better readability to briefly review this work. To represent the human 3D pose by its joints let $\mathbf{X}$ denote the matrix corresponding to P kinematic joints in the 3D space namely $\mathbf{X} = [\mathbf{X}_1 ... \mathbf{X}_P] \in \mathcal{X} \subset \mathbb{R}^{3\times P}$ where $\mathcal{X}$ denotes the space of valid human poses. Akhter&Black [1] (similar to [24, 42]) assumed that all of the 2D joints are observed and estimated a single 3D pose by solving the following op-

timization problem:

$$\min_{\omega,s,\mathbf{R}} C_r + C_p + \beta C_l, \tag{1}$$

where, C_r is a measure of fitness between the estimated 2D joints $\hat{\mathbf{x}} \in \mathbb{R}^{2\times P}$ and the projection and translation of estimated 3D pose $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_1 ... \hat{\mathbf{X}}_P] \in \mathbb{R}^{3\times P}$ to the 2D image coordinate system in a weak perspective camera model (orthographic projection) with scaling factor $s \in \mathbb{R}^+$, rotation $\mathbf{R} \in SO(3)$, and translation $\mathbf{t} \in \mathbb{R}^{2\times 1}$, defined as:

$$C_r = \sum_{i=1}^{P} \|\hat{\mathbf{x}}_i - s\mathbf{R}_{1:2}\,\hat{\mathbf{X}}_i + \mathbf{t}\|_2^2, \tag{2}$$

where, $\mathbf{R}_{1:2}$ denotes the first two rows of the rotation matrix. Note that if the origin of the 3D world coordinate system gets mapped to the origin of the 2D image coordinate system then $\mathbf{t} = \mathbf{0}$; this is usually implemented by centering the 2D and 3D poses. Authors used a sparse representation of the 3D poses similar to [24] where the 3D pose is represented by a sparse linear combination of bases selected using the Orthogonal Matching Pursuit (OMP) algorithm [18] from an overcomplete dictionary of pose atoms, namely $\hat{\mathbf{X}} = \mu + \sum_{i\in\mathcal{I}^*} \omega_i \mathbf{D}_i$, where μ is the mean pose obtained by averaging poses from the CMU motion capture dataset [11] and $\mathcal{I}^*$ denotes the indices of selected bases using OMP with weights ω_i. An overcomplete dictionary of bases was built by concatenating PCA bases from poses of different action classes in the CMU dataset after bone length normalization and Procrustes aligned. The second term C_p in equation (1) is equal to zero if the estimated pose $\hat{\mathbf{X}}$ has valid joint angles for limbs and infinity otherwise. According to the pose-conditioned constraints in [1] a pose has valid joint angles if the upper arms/legs' joint angles map to a 1 in the corresponding occupancy matrix (learned from the ABCD dataset) and the lower arms/legs satisfy two conditions that prevent these bones from bending beyond feasible joint-angle limits (inequalities (4) and (5)). The term C_l in equation (1) penalizes the difference between the squares of the estimated i^{th} bone length l_i and the normalized mean bone length $\bar{l}_i$ *i.e.*, $C_l = \sum_{i=1}^{N} |l_i^2 - \bar{l}_i^2|$ (normalized mean bones calculated from the CMU dataset) with weight β. Note that [1] does not introduce any generative pose model.

As we mentioned earlier, 3D pose estimation from 2D landmark points in monocular RGB images is inherently an ill-posed problem because of losing the depth information. There can be multiple valid 3D poses with similar 2D projection even if all of the 2D joints are observed (see Figure 1). The uncertainty and number of possible valid poses can further increase if some of the joints are missing. The missing joints scenario is more realistic because it happens when either these joints exist in the image but are not confidently detected, due to occlusion and clutter, or do not exist within the borders of the image *e.g.* when only the upper body is visible similar to images from the FLIC dataset [27]. It is observed that thresholding the confidence score obtained from some deep 2D joint detectors (*e.g.* [19, 21, 14]) can be reasonably used as an indicator for the confident detection of a joint. Figure 3 shows the the output of "Stacked Hourglass" 2D joint detector [19] in the absence and presence of a table occluder segmented out from the Pascal VOC dataset [13] and pasted on the left hand of the human subject. On the right-hand-side of each image is shown the heatmap for each joint. It can be seen that the level of the two heatmaps corresponding to the left elbow and left wrist drop after placing the table occluder on the left hand. Newell et al. [19] used the heatmap mean as a confidence measure for detection and threshold it at 0.002 to determine visibility of a joint. Obviously, invisibility of some joints in the image can result in multiple hallucinations for the 2D/3D locations of the joints. Let S_o and S_m denote the set of observed and missing joints, respectively. We have $S_\text{o} \cap S_\text{m} = \emptyset$ and $S_\text{o} \cup S_\text{m} = \{1, 2, ..., P\}$, and let $\alpha = \{\alpha_i\}_{i\in S_\text{o}}$ denote a set of normalized joint scores from the 2D joint detectors such that $\frac{1}{|S_\text{o}|}\sum_{i\in S_\text{o}} \alpha_i = 1$. The missing joints are detected by comparing the confidence score of 2D joint detector with a threshold (0.002 in the case of using Hourglass). For the case of missing joints, we modify the fitness measure to:

$$C_r = \sum_{i\in S_\text{o}} \alpha_i \|\hat{\mathbf{x}}_i - s\mathbf{R}_{1:2}\,\hat{\mathbf{X}}_i + \mathbf{t}\|_2^2. \tag{3}$$

The scores are normalized because they have to be in a comparable range with respect to the C_l term in equation (1) otherwise either C_r is suppressed/ignored in the case of very small confidence scores or the same happens to C_l in the case of very large scores. For example, if the mean of heatmaps from the Hourglass joint detector are directly (without normalization) used as scores the C_r term will be drastically suppressed since the heatmaps are full of close-to-zero values. Note that the optimization problem in equation (1) with the updated C_r term according to equation (3) still outputs a full 3D pose even under missing joints scenario because the 3D pose is constructed by a linear combination of full body basis. However, there is no reason that the output 3D pose should have a close to correct 2D projection due to the missing joint ambiguity added to the depth ambiguity. Optimizing C_r is a non-convex optimization problem over the 3D pose and projection matrix. To obtain an estimate of the 3D torso and projection matrix, we tried both iterating between optimizing over the projection matrix and 3D pose used in [1] as well as the convex relaxation method in [42] as will be presented in the experimental results section. Note that the torso pose variations are much fewer than the full-body. The torso plane is usually vertical and not as flexible as the full body. Hence, it is

much easier to robustly estimate its 3D pose and the corresponding camera parameters.

To generate multiple diverse 3D pose hypotheses consistent with the output of 2D joint detector, we cluster samples from a conditional distribution given the collected 2D evidence. For this purpose, we follow a rejection sampling strategy. Before discussing conditional sampling in subsection 2.2 we describe unconditional sampling as follows.

2.1. Unconditional Sampling

Given the rigidity of human torso compared to the limbs (hands/legs), the joints corresponding to the torso including thorax, left/right hips, and left/right shoulders can be represented using a small size dictionary after an affine transformation/normalization. Given the torso, the upper arms/legs and head are anatomically restricted to be within certain angular limits. The plausible angular regions for the upper arms/legs and head can be represented using an occupancy matrix [1]. This occupancy matrix is a binary matrix that assigns 1 to a discretized azimuthal θ and polar ϕ angle if these angles are anatomically plausible and 0 otherwise. These angular positions are calculated in the local Cartesian coordinate system whose two axis are the "backbone" vector and either the "right shoulder $\rightarrow$ left shoulder" vector (for the upper arms and head) or the "right hip $\rightarrow$ left hip" vector (for the upper hips). Hence, to generate samples for the upper arms/legs and head we just need to take samples from the occupancy matrix at places where the value is 1 and get the corresponding azimuthal and polar angles. Given the azimuthal and polar angles of the head we just need to travel in this direction for the length of the head; we do the same for the length of upper arms and legs to reach the elbows and knees, respectively. The normalized length of the bones is sampled from a Beta distribution with limited range under the constraint that similar bones have similar length *e.g.* both upper arms have the same length.

According to [1], the lower arm/leg bone $\mathbf{b}_{p_1 \rightarrow p_2} = \mathbf{X}_{p_2} - \mathbf{X}_{p_1}$, where p_2 and p_1 respectively correspond to either "wrist and elbow" or "ankle and knee" is at a plausible angle if it satisfies two constraints. The first constraint is:

$$\mathbf{b}^\top \mathbf{n} + d < 0, \tag{4}$$

where $\mathbf{n}$ and d are functions of the azimuthal θ and polar ϕ angles of their parent bone namely the upper arm or leg (resulting in pose-dependent joint angle limits) learned from the ABCD dataset. The above inequality defines a separating plane, with normal vector $\mathbf{n}$ and distance from origin d, that attempts to prevent the wrist and ankle from bending in a direction that is anatomically impossible. Obviously, for a very negative offset vector d this constrain is always satisfied. Therefore, during learning of $\mathbf{n}$ and d the second norm of d is minimized, namely $\min_{\mathbf{n},d} \|d\|_2$ s.t. $\mathbf{B}^\top \mathbf{n} < -d\mathbf{1}$, where $\mathbf{B}$ is a matrix built by column-wise concatenation of all $\mathbf{b}$ instances in the ABCD dataset whose parents are at the same θ and ϕ angular location. The second constraint to satisfy is that the projection of normalized $\mathbf{b}$ (to unit length) onto the separating plane using the orthonormal projection matrix $\mathbf{T} = [\mathbf{T}_1; \mathbf{T}_2; \mathbf{T}_3]$, whose first row $\mathbf{T}_1$ is along $\mathbf{n}$, has to fall inside a bounding box with bounds $[bnd_1, bnd_2]$ and $[bnd_3, bnd_4]$, namely:

$$\begin{aligned} bnd_1 \leq \mathbf{T}_2\mathbf{b}/\|\mathbf{b}\|_2 \leq bnd_2, \\ bnd_3 \leq \mathbf{T}_3\mathbf{b}/\|\mathbf{b}\|_2 \leq bnd_4, \end{aligned} \tag{5}$$

where, bounds bnd_1, bnd_2, bnd_3, and bnd_4 are also learned from the ABCD dataset. To generate a sample $\mathbf{b}$ that satisfies the above constraints, we first generate two random values $u_2 \in [bnd_1, bnd_2]$ and $u_3 \in [bnd_3, bnd_4]$ and set $u_1 = (\max(1 - u_2^2 - u_3^2, 0))^{1/2}$. We then generate two candidates $\mathbf{u}^\pm = (\pm u_1, u_2, u_3)/\|(u_1, u_2, u_3)\|_2$ from which only one can be on the valid side of the separating plane satisfying inequality (4). To check, we first undo the projection and normalization by $\mathbf{b}^\pm = l\,\mathbf{T}^{-1}\mathbf{u}^\pm$, where l is a sample from the bone length distribution on $\mathbf{b}$. A sample "$\mathbf{b}$" is accepted only if it satisfies inequality (4). Note that similar bones have the same length therefore we sample their length only once for each pose. The prior model can be written as below according to a Bayesian graph on the kinematic chain:

$$\begin{aligned} &p(\mathbf{X}) = p(\mathbf{X}_{i\in\text{torso}})p(\mathbf{X}_{\text{head}}|\mathbf{X}_{i\in\text{torso}})\times \\ &p(\mathbf{X}_{i\in\text{ l/r elbow}}|\mathbf{X}_{i\in\text{torso}})p(\mathbf{X}_{i\in\text{ l/r wrist}}|\mathbf{X}_{i\in\text{ l/r elbow}}, \mathbf{X}_{i\in\text{torso}})\times \\ &p(\mathbf{X}_{i\in\text{ l/r knee}}|\mathbf{X}_{i\in\text{torso}})p(\mathbf{X}_{i\in\text{ l/r ankle}}|\mathbf{X}_{i\in\text{ l/r knee}}, \mathbf{X}_{i\in\text{torso}}), \end{aligned} \tag{6}$$

where $p(\mathbf{X}_{i\in\text{torso}})$ is the probability of selecting a torso from the torso dictionary which we assumed is uniform. The torso joints $\mathbf{X}_{i\in\text{torso}}$ are used to determine the local coordinate system for the rest of the joints. We have removed torso joints in the equations below for notational convenience. We have:

$$p(\mathbf{X}_i) = \frac{1}{l_{\text{bone}}^2 |\sin(\phi_i)|} p(l_{\text{bone}})\, p(\theta_i, \phi_i), \tag{7}$$

for (i, bone) being from (l/r knee, upper leg) , (head, neck + head bone), or (l/r elbow, upper arm). The multiplier factor in (7), which is the inverse of Jacobian determinant for a transformation from the Cartesian to spherical coordinate system, is to ensure that the left side sums up to one if $\int_l \int_\theta \int_\phi p(l)\, p(\theta, \phi) d\phi\, d\theta\, dl = 1$, since $dx\, dy\, dz = l^2 |\sin(\phi)|\, dl\, d\theta\, d\phi$. For lower limbs we have:

$$p(\mathbf{X}_i|\mathbf{X}_{pa(i)}) \propto p(l_{\text{bone}})\mathbf{1}_{valid}(\mathbf{X}_i, \mathbf{X}_{pa(i)}) \tag{8}$$

where $(i, pa(i), \text{bone})$ is from (l/r wrist, l/r elbow, forearm) or (l/r ankle, l/r knee, lower leg), and $\mathbf{1}_{valid}(\mathbf{X}_i, \mathbf{X}_{pa(i)})$ is an indicator function that nulls the probability of configurations whose angles does not satisfy the constraints in inequalities (4) and (5) for $\mathbf{b} = \mathbf{X}_i - \mathbf{X}_{pa(i)}$. Conditional

sampling is carried out by rejection sampling discussed in the next subsection.

2.2. Conditional Sampling

We run a 2D joint detector on the input image I and get an estimate of the 2D joint locations $\hat{\mathbf{x}}$ with confidence scores α. Then, to obtain a reasonable estimate of torso $\hat{\mathbf{X}}_{i\in\text{torso}}$ and camera parameters namely $(\hat{\mathbf{R}}, \hat{\mathbf{t}}, \hat{s})$, we run a 2D-to-3D pose estimator capable of handling missing joints (we modified [1] and [42] to handle missing joints; see equation (3)). Note that we are not restricted to any particular 2D/3D pose estimator and any 2D joint detector that estimates 2D joint locations $\hat{\mathbf{x}}$ and their confidence scores α and any 2D-to-3D pose estimator can be used in the initial stage. We then assume that the estimated camera parameters and $\hat{\mathbf{X}}_{i\in\text{torso}}$ are reasonably well estimated and keep them fixed. Note that the human torso and its pose (usually vertical) does not vary much compared to the whole body pose. We do not include the estimated camera parameters and 3D torso in our formulation below for notational convenience. From the Bayes rule we have:

$$p(\mathbf{X}|\hat{\mathbf{x}}, \alpha) \propto p(\mathbf{X})\, p(\hat{\mathbf{x}}, \alpha|\mathbf{X}). \tag{9}$$

We define:

$$p(\hat{\mathbf{x}}, \alpha|\mathbf{X}) \propto \prod_{i\in \text{ limb}\cap S_\text{o}} \mathbf{1}(\|\hat{\mathbf{x}}_i - \hat{s}\,\hat{\mathbf{R}}_{1:2}\,\mathbf{X}_i + \hat{\mathbf{t}}\|_2 < \tau_i)$$

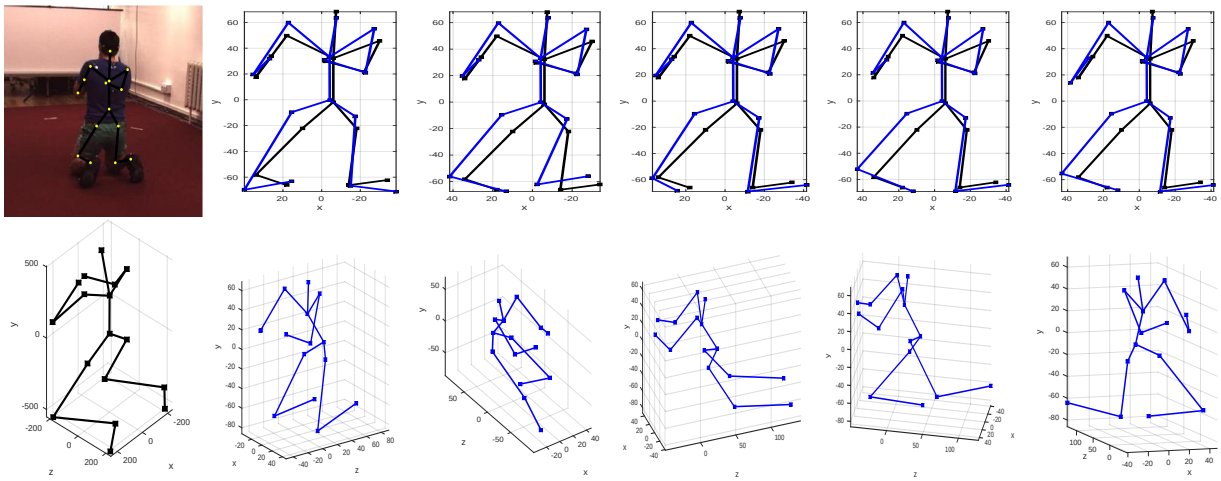
where $\mathbf{1}(.)$ is the indicator function depending on the 2D distance between detected joints and the projected 3D pose under an acceptance threshold defined by $\tau_i = 0.25\,\hat{s}\,\bar{l}_\text{limb}/\alpha_i$, where $\bar{l}_\text{limb}$ is the mean limb length, $\hat{s}$ is the estimated scaling factor, α_i is the i^th joint normalized confidence score, and the factor 0.25 was chosen empirically. The likelihood function defined above accepts prior (unconditional) samples $\mathbf{X}^{(q)} \sim p(\mathbf{X})$ whose projected joints to the image coordinate system are within a distance not greater than thresholds τ_i from detected limb joints. The inverse proportion of the threshold to the confidence α_i allows acceptance in a larger area if the confidence score is smaller for the i^th limb joint and therefore considering the 2D joint detection uncertainty. Note that there is no indicator function in the likelihood function for the missing limb joints which allows acceptance of all anatomically plausible samples for limb joints from S_m. Note that even though torso pose estimation is a much easier problem compared to the full body pose estimation, a poorly estimated torso, *e.g.* due to occlusion, can adversely affect the quality of conditional 3D pose samples.

2.3. Generating Diverse Hypotheses

The diversification is implemented in two stages: (I) we sampled the occupancy matrix at 15 equidistant azimuth and 15 equidistant polar angles for the upper limbs and accept the samples if the occupancy matrix had a 1 at these locations. For the lower limbs, we sampled 5 equidistant points along each u_2 and u_3 directions between $[bnd_1, bnd_2]$ and $[bnd_3, bnd_4]$, respectively. (II) To generate fewer number of pose hypothesis, we use the kmeans++ algorithm [3] to cluster the posterior samples into a desired number of diverse clusters and take the nearest neighbor 3D pose sample to each centroid as one hypothesis. Kmeans++ operates the same as Kmeans clustering except that it uses a diverse initialization method to help with diversification of final clusters. Note that we cannot take the centroids as hypotheses since there is no guarantee that the mean of 3D poses is still a valid 3D pose. Figure 4 shows five hypotheses given the output of Hourglass 2D joint detector for the top-left image and detections shown by yellow points. In Figure 4, the 2D detection of joints are shown by the black skeleton and the diversified hypotheses that are consistent with the 2D detections are shown by the blue skeletons. It can be seen that even though the 2D projection of these pose hypotheses are very similar, they are quite different in 3D. To generate the pose hypotheses in Figure 4, we estimated the 3D torso and projection matrix using [1]. s

3. Experimental Results

We empirically evaluated the proposed "multi-pose hypotheses" approach on the recently published Human3.6M dataset [15]. For evaluation, we used images from all 4 cameras and all 15 actions associated with 7 subjects for whom ground-truth 3D poses were provided namely subjects S1, S5, S6, S7, S8, S9, and S11. The original videos (50 fps) were downsampled (in order to reduce the correlation of consecutive frames) to built a dataset of 26385 images. For further evaluation, we also built two rotation datasets by rotating H36M images by 30 and 60 degrees. We evaluated the performance by the mean per joint error (millimeter) in 3D by comparing the reconstructed pose hypotheses against the ground truth. The error was calculated up to a similarity transformation obtained by Procrustes alignment. The results are summarized in Table 1 for various methods and actions. For a fair comparison, the limb length of the reconstructed poses from all methods were scaled to match the limb length of the ground-truth pose. The bone length matching obviously lowers the mean joint errors but makes no difference in our comparisons. One can see that the best (lowest Euclidean distance from the ground-truth pose) out of only 5 generated hypotheses by using [1] as baseline for 3D torso and projection matrix estimation is considerably better than the single 3D pose output by [1] for all actions. We also used the 2D-to-3D pose estimator by Zhou et al. [42] with convex-relaxation as baseline and observed considerable improvement compared to [1] in both 3D pose and projection matrix estimation. Using [42] as baseline to estimate the 3D torso and



(a) (b)

Figure 4. (a): The input image and the corresponding 3D pose. (b): Generation of five diverse 3D pose hypotheses consistent with the 2D joint detections.

Method	Directions	Discussion	Eating	Greeting	Phoning	Posing	Purchases	Sitting	SitDown
Ours (No KM++/[42])	**63.12**	**55.91**	**58.11**	**64.48**	**68.69**	**61.27**	**55.57**	**86.06**	**117.57**
Ours (k=20/[42])	77.08	71.15	75.39	79.01	84.68	74.90	72.37	102.17	131.46
Ours (k=5/[42])	82.86	77.52	81.60	85.20	90.93	80.46	78.75	109.27	138.71
Zhou et al. [42]	80.51	74.56	73.95	85.43	88.96	82.02	76.21	107.43	146.47
Ours (k=5/[1])	**105.14**	**100.28**	**107.75**	**106.88**	**111.44**	**105.74**	**101.18**	**124.87**	**147.48**
Akhter&Black [1]	133.80	128.03	124.47	133.47	133.93	136.63	128.30	133.61	162.01
Chen et al. [8]	145.37	139.11	140.24	149.13	149.61	154.30	147.04	161.49	200.06

	Smoking	TakingPhoto	Waiting	Walking	WalkingDog	WalkTogether	Average
Ours (No KM++/[42])	**71.02**	**71.21**	**66.29**	**57.07**	**62.50**	**61.02**	**67.99**
Ours (k=20/[42])	85.90	84.49	80.41	71.57	78.41	74.92	82.93
Ours (k=5/[42])	91.79	90.06	86.43	77.93	85.45	81.49	89.23
Zhou et al. [42]	90.61	93.43	85.71	80.03	90.89	85.73	89.46
Ours (k=5/[1])	**113.61**	**105.58**	**105.80**	**100.28**	**106.25**	**104.63**	**109.79**
Akhter&Black [1]	135.75	132.92	133.93	133.84	131.77	134.80	134.48
Chen et al. [8]	152.37	159.18	152.67	148.20	156.10	147.71	153.51

Table 1. Quantitative comparison on the Human3.6M dataset evaluated in 3D by mean per joint error (mm) for all actions and subjects whose ground-truth 3D poses were provided.

projection matrix we generated multiple 3D pose hypotheses. Since the accuracy of [42] is already high, the best out of 5 pose hypotheses cannot significantly lower the average joint distance from the single 3D pose output by [42]. However, by increasing the number of hypotheses we started to observe improvement. Table 1 also includes the best hypothesis out of conditional samples from only the first diversification stage *i.e.*, by diversifying conditional samples and using no kmeans++ clustering (shown by No KM++), using [42] as base. This achieves the lowest joint error in comparison to other baselines. The pose hypotheses can be generated very quickly ($<$ 2 seconds) in Matlab on an Intel i7-4790K processor.

We also used Deep3D of Chen et al. [8] as another baseline. The Deep3D [8] is a 3D pose estimator that directly regresses to the 3D joint locations directly from a monocular RGB input image. Deep3D had the highest mean joint errors as shown in Table 1. We also observed that the pre-trained Deep3D is very sensitive to image rotation and usually outputs an anatomically implausible 3D pose if the input image is rotated. But other 2D-to-3D pose estimation baselines which decouple the projection matrix and the 3D pose are quite robust to rotation of the input image. Figure 5 shows the Percentage of Correct Keypoints (PCK) versus an acceptance distance threshold in millimeter for various baselines and H36M dataset variations namely the original H36M and 30/60 degree rotations. One can see that the PCK of Deep3D drops drastically by rotating the input image. This is partly due to insufficient number of tilted samples in the training set (H36M plus synthetic images). One of the main problems of purely discriminative approaches such as [8] is their extreme sensitivity to data manipulation. On the other hand, humans can learn from a few examples and still not suppress the rarely seen cases compared to the frequently seen ones.

In a realistic scenario with occlusion, the location of

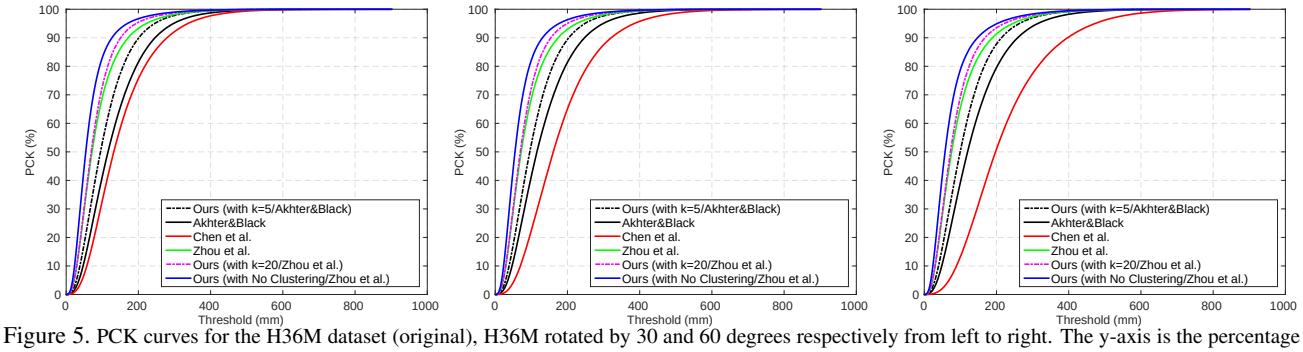


Figure 5. PCK curves for the H36M dataset (original), H36M rotated by 30 and 60 degrees respectively from left to right. The y-axis is the percentage of correctly detected joints in 3D for a given distance threshold in millimeter (x-axis).

Method	Directions	Discussion	Eating	Greeting	Phoning	Posing	Purchases	Sitting	SitDown
Ours (k=5/[1])	**98.44**	**93.70**	**102.62**	**97.50**	**96.29**	**98.90**	**93.32**	**105.51**	**110.07**
Akhter&Black [1]	118.02	112.55	111.27	117.46	111.77	122.27	112.23	107.27	126.95
Ours (k=5/[1])	**108.60**	**105.85**	**105.63**	**109.01**	**105.47**	**109.93**	**102.01**	**111.25**	**119.57**
Akhter&Black [1]	153.80	149.14	135.44	155.06	139.62	156.46	149.05	126.33	141.89
Ours (k=5/[1])	**125.03**	**121.77**	**115.13**	**124.11**	**116.92**	**123.75**	**116.42**	**119.63**	**130.81**
Akhter&Black [1]	185.57	180.43	158.55	185.65	162.39	185.78	178.81	145.15	155.29

	Smoking	TakingPhoto	Waiting	Walking	WalkingDog	WalkTogether	Average	Average Diff.
Ours (k=5/[1])	**97.53**	**97.63**	**99.43**	**90.23**	**97.27**	**95.21**	**98.24**	
Akhter&Black [1]	113.22	120.61	119.97	115.81	116.60	115.62	116.11	17.87
Ours (k=5/[1])	**107.76**	**107.05**	**111.34**	**108.38**	**106.96**	**110.28**	**108.61**	
Akhter&Black [1]	142.98	152.65	155.27	155.18	151.88	155.00	147.98	39.37
Ours (k=5/[1])	**120.60**	**118.38**	**127.13**	**125.89**	**121.61**	**127.62**	**122.32**	
Akhter&Black [1]	165.47	177.44	186.20	189.66	183.01	186.25	175.04	**52.72**

Table 2. Quantitative comparison on the Human3.6M dataset when 0 (top pair), 1 (middle pair), and 2 (bottom pair) limb joints are missing.

some 2D joints cannot be accurately detected. The added uncertainty caused by occlusion makes one expect a larger average estimation error for the estimated 3D pose from a single-output pose estimator compared to the best 3D pose hypothesis. To test this, we ran experiments with different number of missing joints (0, 1 and 2) selected randomly from the limb joints including l/r elbow, l/r wrist, l/r knee, and l/r ankle. Table 2 shows the mean per joint errors for the 3D pose estimated by the modified version of Akhter&Black [1] that can handle missing joints compared to the best out of five hypotheses generated by our method when 0, 1, and 2 limb joints are missing. In this test, we used the ground-truth 2D location of the joints and randomly selected the missing joints. One can see that by increasing the number of missing joints the performance gap between the estimated 3D pose and the best 3D pose hypothesis increases. This underscores the importance of having multiple hypothesis for more realistic scenarios.

4. Conclusion

There usually exist multiple 3D poses consistent with the 2D location of joints because of losing the depth information in monocular images. The uncertainty in 3D pose estimation increases in the presence of occlusion and imperfect 2D detection of joints. In this paper, we proposed a way to generate multiple valid and diverse 3D pose hypotheses consistent with the 2D joint detections. These pose hypotheses can be ranked later by more detailed investigation of the image beyond the 2D joint locations or based on some contextual information. To generate these pose hypotheses we used a novel unbiased generative model that only enforces pose-conditioned anatomical constraints on the joint-angle limits and limb length ratios. This was motivated by the pose-conditioned joint limits from [1] after identifying bias in typical MoCap datasets. Our compositional generative model uniformly spans the full variability of human 3D pose which helps in generating more diverse hypotheses. We performed empirical evaluation on the H36M dataset and achieved lower mean joint errors for the best pose hypothesis compared to the estimated pose by other recent baselines. The 3D pose output by the baseline methods could also be included as one hypothesis but to investigate our hypothesis generation approach we did not do so in the experimental results. Our experiments show the importance of having multiple 3D pose hypotheses given only the 2D location of joints especially when some of the joints are missing. We hope our idea of generating multiple pose hypotheses inspire a new line of future work in 3D pose estimation considering various ambiguity sources.

References

[1] I. Akhter and M. J. Black. Pose-conditioned joint angle limits for 3D human pose reconstruction. In *CVPR*, pages 1446–1455, June 2015. 2, 3, 4, 5, 6, 7, 8

[2] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele. Multi-view pictorial structures for 3d human pose estimation. In *British Machine Vision Conference (BMVC)*, September 2013. 3

[3] D. Arthur and S. Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. 6

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

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3d pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(10):1929–1942, 2016. 3

[6] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016. 3

[7] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *CVPR*, pages 3618–3625, 2013. 3

[8] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV)*, 2016. 3, 7

[9] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014. 3

[10] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *CVPR*, 2016. 3

[11] P. Doe. Cmu human motion capture database. availabel online at:, 2003. 2, 4

[12] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99:190–214, 2012. 3

[13] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan. 2015. 4

[14] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, May 2016. 4

[15] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 2, 6

[16] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *CVPR*, 2004. 3

[17] A. M. Lehrmann, P. V. Gehler, and S. Nowozin. A non-parametric bayesian network prior of human pose. In *CVPR*, pages 1281–1288, 2013. 3

[18] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, pages 3397–3415, Dec. 1993. 4

[19] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, May 2016. 2, 3, 4

[20] D. Park and D. Ramanan. N-best maximal decoders for part models. In *ICCV*, 2011. 3

[21] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, June 2016. 4

[22] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taix, M. Mller, H.-P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV*, 2011. 3

[23] G. Pons-Moll, D. J. Fleet, and B. Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, pages 2345–2352, June 2014. 3

[24] V. Ramakrishna, T. Kanade, and Y. Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*, 2012. 3, 4

[25] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3d pose estimation in the wild. 2016. 3

[26] G. Rogez, P. Weinzaepfel, and C. Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 3

[27] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *CVPR*, pages 3674–3681, 2013. 3, 4

[28] L. Sigal, A. Balan, and M. J. Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision*, 87:4–27, 2010. 2

[29] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *CVPR*, June 2004. 2

[30] L. Sigal, M. Isard, H. Haussecker, and M. J. Black. Loose-limbed people: Estimating 3D human pose and motion using non-parametric belief propagation. *International Journal of Computer Vision*, 98(1):15–48, May 2011. 3

[31] 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 *CVPR*, pages 3634–3641, 2013. 3

[32] E. Simo-Serra, A. Ramisa, G. Alenyà, C. Torras, and F. Moreno-Noguer. Single image 3d human pose estimation from noisy observations. In *CVPR*, 2012. 3

[33] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3d human tracking. In *CVPR*, 2003. 3

[34] 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 *Advances in neural information processing systems*, pages 1799–1807, 2014. 3

[35] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 3

[36] L. van der Maaten and G. E. Hinton. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 2

[37] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, 2014. 3

[38] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, June 2016. 3

[39] H. L. X. W. Xiao Chu, Wanli Ouyang. Crf-cnn: Modeling structured information in human pose estimation. In *NIPS*, 2016. 3

[40] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016. 3

[41] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 3

[42] X. Zhou, S. Leonardos, X. Hu, and K. Daniilidis. 3d shape estimation from 2d landmarks: A convex relaxation approach. In *CVPR*, pages 4447–4455, June 2015. 2, 3, 4, 6, 7

[43] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, June 2016. 3