Robust Monocular 3D Human Motion with Lasso-Based Differential Kinematics

Abed Malti
Laboratoire de Génie Biomédical
Université de Tlemcen, Algeria
abed.malti@gmail.com

Abstract

This work introduces a method to robustly reconstruct 3D human motion from the motion of 2D skeletal landmarks. We propose to use a lasso (least absolute shrinkage and selection operator) optimization framework where the ℓ_1-norm is computed over the vector of differential angular kinematics and the ℓ_2-norm is computed over the differential 2D reprojection error. The ℓ_1-norm term allows us to model sparse kinematic angular motion. The minimization of the reprojection error allows us to assume a bounded noise in both the kinematic model and the 2D landmark detection. This bound is controlled by a scale factor associated to the ℓ_2-norm data term. A posteriori verification condition is provided to check whether or not the lasso formulation has allowed us to recover the ground-truth 3D human motion. Results on publicly available data demonstrates the effectiveness of the proposed approach on state-of-the-art methods. It shows that both sparsity and bounded noise assumptions encoded in lasso formulation are robust priors to safely recover 3D human motion.

1. Introduction

Solving 3D skeletal human motion recovery from 2D landmark motion is challenging for many critical applications using monocular video: crowd surveillance [10], try-on clothing [14,23], mixed reality [22], human action recognition [48], sports [8] and social-media applications [26]. The main challenge comes from the fact that this problem is ill-posed since more than one combination of camera motion and 3D human motion can produce similar 2D motion [34, 40]. Since the past decade, 3D human from single image has become a proper field of investigation and has received a particular attention with several methodological contributions [9,28,43,56]. One of the most popular way of solving this problem is to consider the pose to be retrieved as a linear combination of a pre-trained dictionary [17,43]. An optimization framework is used to estimate the coefficients of the linear combination and the rigid camera pose. This approach is non-convex and has inherent wrong reconstructions due to the possibility of using a wrong initial pose. 3D humans from a single image are usually extended to 3D humans from image sequence by constraining smooth variations over the coefficients of the basis shapes and the camera motion [53, 60]. However, such methods may fail if a subset of 3D camera and body poses are not accurate. In this work, we use a prior on separate and sparse motion to constrain frame-to-frame 3D reconstruction (see figure 1). Instead of using a dictionary of shapes, some authors used articulated skeleton as in [50], [20], [31], probabilistic graphical models as in [47], [3], explicit regression by [16], [1]. Most of these methods are based on ℓ_2-norm minimization which is strongly affected by noise in data and skeleton modeling. In this work, we consider an ℓ_2-norm minimization on the reprojection error which relaxes equality constraint to account for noise and assumes an upper bound on the noise affecting 2D detection and skeleton modeling [18]. 3D human motion from 2D sequence of skeleton landmarks has been initiated in [6]. A least squares method was used to recover the motion of the articulated angles and the rigid motion of the camera. Park et al. [40] proposed to use constant limb length through time as spatial and temporal constraints to smooth the 3D articulated motion. They showed that at every reconstruction, there exist two solutions which satisfy each instantaneous 2D projection and articulation constraint. Many other works on Non-Rigid Structure-from-Motion with application to human 3D motion recovery have followed [2, 15, 62]. Specific skeleton-based works using video sequences have also been attempted [37, 45, 54]. Recently, [34] proposed a criterion combining kinematic and projection matrices to evaluate a posteriori if the recovered 3D human motion corresponds to the ground-truth. Our contribution extends this method by considering noise in landmark detection and skeleton modeling. We propose a lasso formulation that allows us to account for both this noise and for the sparsity in human motion. We use SMPL [32] a parametric kinematic skele-

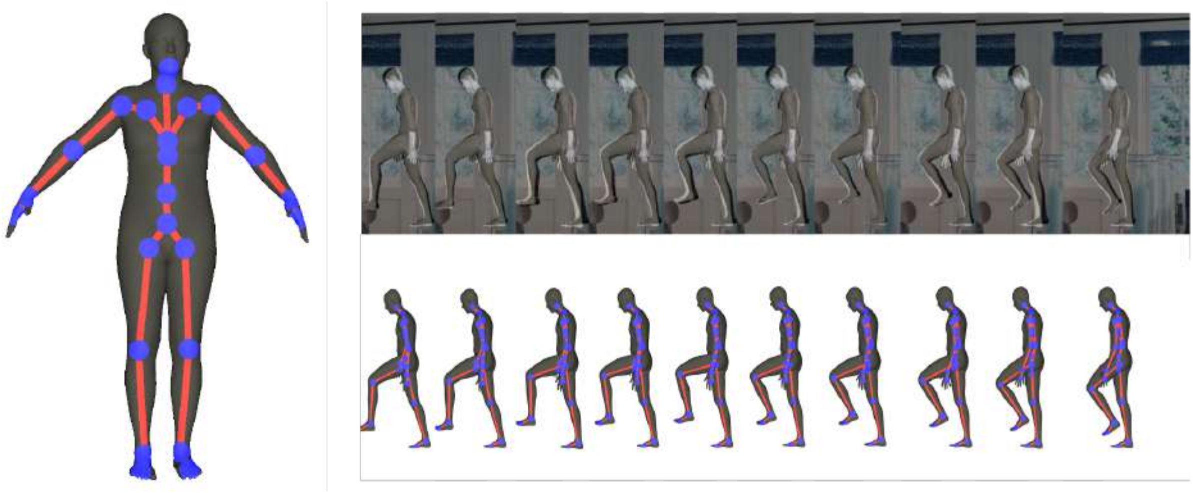

Figure 1. Left: SMPL model with the associated 24 × 3-angular joints [32]. This skeleton has 72 rotational degress of freedom and 3 translational degrees of freedom. Top-right: First 10 frames from test sequence *run0/27_09_c0003* in SURREAL dataset. Bottom-right: 3D motion reconstruction with proposed method. From right-to-left frames, the right hip then the right knee are the most prominent joint motion of this sequence.

ton to intrinsically constrain the reconstructed 3D motion. This model is represented by a set of angular articulations connected by limbs (bones). Each single rotation represents a Degree of Freedom (DoF). Every joint articulation has 3 degrees of freedom (see figure 1-Left). The total number of angular degrees of freedom gathers 72 angles including the orientation of the root hip joint with respect to the camera view. Three more translational degrees of freedom complete the parametric description of the human posture. From a frame to a next one, the ℓ_1-norm allows us to denoise the articulated motion by sending to zero angular motions that are due to jitter in 2D landmark motion. We assume known an upper bound estimate of the noise in skeleton modeling and landmark detection. The proposed formulation allows us to recover 3D motion even if the number of detected landmarks is smaller than the number of DoFs. We assume that the 2D landmarks can be detected with a state-of-the-art 2D detection of human pose [7, 24].

Contributions and specificities. In this paper, we propose a lasso-based formulation of the 3D human motion recovery problem. We minimize both the ℓ_1-norm of the skeleton's degrees of freedom and the ℓ_2-norm of the reprojection error from detected 2D landmarks. The ℓ_1-norm term encodes the denoising and motion sparsity. The ℓ_2-norm term relaxes the equality constraint on the 2D reprojection to account for noise in landmarkd detection and skeleton's modeling. We derive sufficient conditions to verify a posteriori that the recovered 3D motion corresponds to groundtruth in the noise-free case. We ran experiments to show the robustness of the proposed method to recover accurate support of the sparse angular motion and accurate angular motion. We compare and validate the proposed method to 4 state-of-the-art methods on 3 publicly available dataset.

2. Related Work

It is difficult to cover the overall state-of-the-art in 3D human recovery from monocular views. Our study tends to reveal the current methodologies that prevails in the community. In this section, we subdivide related approaches into two sub-categories: single image based and sequential image based approaches. The proposed approach is part of the sequential image based category since it uses the 2D motion between current and previous frame to infer the 3D human motion.

Single Image based. [11, 29, 35, 51] used CNNs (Convolutional Neural Networks) to regress 3D human pose from 2D joint locations, semantics or even raw images. Most of these proposed neural networks are difficult to train and the outcome is very dependent to hyperparameter training (batch size, learning rate, type of optimizer, etc). In this paper, we use only kinematic based priors without training any network. [36, 38, 42, 44] used a convolutional neural network to infer 3D articulated pose from an image without using 2D joint locations as input. In this work, we consider the 2D joint location as available and we take into account the noise that is inherent to any landmark detection procedure. [37, 45, 54, 59] combined neural network regression with skeleton kinematic to predict 2D and 3D joint locations. [33, 61] proposed an approach that combines factorization approaches with CNNs approaches. They designed a neural network to predict the coefficients of canonical 3D human pose and camera viewpoint in two separate channels. These end-to-end methods seem to have an order of magnitude in the 3D reconstruction quality however their heavy computational process is a bottelneck for their application to 3D reconstructions from image sequence. [46] used pose sampling manifold to remove ambiguities from 3D reconstruction regressed by pose networks. Such method which relies on low dimensional parametric dense body are difficult to train and requires many hyper-parameter tuning. It is also difficult to obtain 3D reconstructions aligned with the corresponding 2D landmarks because neural networks fail in regressing accurate translation poses.

Sequential Image based. Factorization methods on image sequences were widely attempted [15, 55, 62]. A natural and intuitive representation of a human body is an articulated kinematic structure. This model allows us to constrain the motion with specific and appropriate degrees of freedom through a kinematic chain representation. [21, 40] propose to reconstruct a 3D articulated trajectory given the trajectories of 2D landmarks. These methods use reprojection costs based on the ℓ_2-norm distance that is known to be non-robust to noise in 2D landmark detection. In this work, we propose to use an ℓ_1-norm together with an ℓ_2-norm in the minimization framework. Recently, [34] proposed a sparse articulated 3D motion recovery using an ℓ_1-norm. However, this approach does not take into account noise in 2D detection and consider the whole degrees of freedom as one single vector. The present work solves these issues by considering both noise and sparsity in human motion. As will be shown, such approach ensures a better recovery of the limbs that move from one frame to another. [37, 53] used the image sequence consistency to impose a temporal bone lengths constraints for both periodic and non-periodic human movements. Constraining bone lengths constancy is not robust since noise in 2D landmark detection can introduce ambiguity in bone lengths and poses recovery. [28, 30, 49, 57] used a neural network architecture to learn latent poses with self-attention kinematics to produce plausible 3D shapes. [4] used bundle-adjustment-based to reconstruct 3D humans with temporal coherence. The advantage of such approach is to resolve ambiguities when having multiple point of views. However, it can run only in an offline mode and requires a large amount of images. Spatial transformers to 3D human reconstruction were also introduced as a new trend in neural networks [58]. Such method cannot be applied online since it requires futur frames to predict middle frame 3D pose. In contrast to methods predicting low dimensional 3D feature descriptors, [13] proposed to regress dense vertices positions by training a neural network that extract features from a single image, which are then used in a gradient descent algorithm. This method lack of physical constraints and can reconstruct non plausible shapes.

Notation. Normal letters = scalars, *e.g.*, $a, b, A, B, etc.$ Bold small letters = vectors, *e.g.*, $\mathbf{a}, \underline{,}etc.$ Bold capital letters = matrices, *e.g.*, $\mathbf{A}, \mathbf{B}, etc.$ Calligraphic letters = sets, *e.g.*, $\mathcal{A}, \mathcal{B}, etc.$ ith element of vector $\mathbf{a}$: a_i. Element located at row i and column j of $\mathbf{A}$: $A(i, j)$. $\mathbf{a}^{\mathrm{T}}$ denotes the transpose of $\mathbf{a}$. $\bar{\mathcal{A}}$ denotes the complementary set of $\mathcal{A}$. If $\mathcal{V}$ is a set of indices, $\mathbf{A}_{\mathcal{V}}$ represents the submatrix of $\mathbf{A}$ made up of the columns indexed by $\mathcal{V}$. $\mathbf{A}_k$ stands for the k-th column of $\mathbf{A}$. $|\mathcal{V}|$ denotes the cardinal of $\mathcal{V}$. $|a|$ stands for the absolute value of real number a. There should be no confusion from the context in using the same symbol for the cardinal of a set and the absolute value of a real number. $\mathbf{a}_{\mathcal{V}}$ represent either the restriction of $\mathbf{a}$ to the indices in $\mathcal{V}$, or the vector which coincides with $\mathbf{a}$ on the indices in $\mathcal{V}$ and is extended to zero outside $\mathcal{V}$. It should be clear from the context which notation is meant. $[\mathbf{A}, \mathbf{B}]$ denotes the row-wise concatenation of matrices $\mathbf{A}$ and $\mathbf{B}$. $[n] = \{1, \ldots, n\}$, $n \in \mathbb{N}$. $[n]^s$: All subsets of $[n]$ of cardinal s. $\operatorname{supp}(\mathbf{a}) = \{i : a(i) \neq 0\}$ is the set of integers indexing the non-zero elements of $\mathbf{a}$. $[\mathbf{a}_i]_{i=1}^k = [\mathbf{a}_1 \ldots \mathbf{a}_k]$. $\{\mathbf{a}_i\}_{i=1}^k = \{\mathbf{a}_1 \ldots \mathbf{a}_k\}$. $\ker(\mathbf{A}) = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$. $\operatorname{span}(\mathbf{A}) = \{\mathbf{y} : \mathbf{A}\mathbf{x} = \mathbf{y}, \mathbf{x} \text{ real vector}\}$. $\|\cdot\|_0$ is the ℓ_0-norm. $\mathbf{A}^+$ denotes the Moore-Penrose pseudo inverse of matrix $\mathbf{A}$. The symbol $[\mathbf{A}, \mathbf{B}]$ stands for horizontal concatenation of matrices. Finally, $\mathbf{1}_n$ is the n-vector where every element is equal to 1. $\mathbf{0}$ is the vector of zeros. Its dimension will be understood from the context.

3. Mathematical definitions and modeling

3.1. Definitions and Assumptions

The human skeleton can be seen as a multi-serial kinematic chains with the hip as a common root joint. Every kinematic joint is composed of a set of angular degrees of freedom (ADoFs). If the total number of ADoFs is n, then a skeleton angular motion can be uniquely represented by a real n-vector of angular motion $\mathbf{\Omega} = [\omega_1, \ldots, \omega_n]$. Let us assume $\mathcal{F} = \{1, \ldots, n\}$ as the set of global indices of vector $\mathbf{\Omega}$.

Definition 1 *The support of the global motion vector* $\mathbf{\Omega}$ *is the subset composed of global indices of its non-zero elements*

$$\mathcal{V} := \{i \in \mathcal{F},\ \textit{s.t.}\ \omega_i \neq 0\}\,. \tag{1}$$

Since $\mathbf{\Omega}$ is composed of zeros and non-zeros elements, its ℓ_1-norm can or cannot be differentiable. The concept of gradient is then extended to subgradient so that,

Definition 2 *The subdifferential of* $\|\mathbf{\Omega}\|_1$ *is called subgradient and is a set defined as:*

$$\partial \|\mathbf{\Omega}\|_1 = \left\{\mathbf{\Sigma} | \mathbf{\Sigma}^{\mathrm{T}} \mathbf{\Omega} = \|\mathbf{\Omega}\|_1, \|\mathbf{\Sigma}\|_\infty \leq 1\right\} \tag{2}$$

$$= \{\mathbf{\Sigma} | \mathbf{\Sigma}_i = \textit{sign}\,(\omega_i)\,,\ \textit{if}\, \omega_i \neq 0 \ \textit{and}\ |\mathbf{\Sigma}_i| \leq 1\, \textit{otherwise}\} \tag{3}$$

Where $\operatorname{sign}(\omega_i) = 1$ if $\omega_i > 0$ and $\operatorname{sign}(\omega_i) = -1$ is $\omega_i < 0$. If $\omega_i = 0$, its subdifferential is not unique and can be any real number between -1 and 1.

3.2. Observation Model with Noise

Let us consider the 2D differential motion vector $\mathbf{y} \in \mathbb{R}^{2l}$ of $l \leq \gamma$ joints from two successive monocular frames. Let us assume $\mathbf{P} \in \mathbb{R}^{2l\times 3l}$, $\mathbf{D} \in \mathbb{R}^{3l\times 6}$ and $\mathbf{J} \in \mathbb{R}^{3l\times n}$ as respectively the Jacobians matrices of perspective projection, camera-to-skeleton rigid pose and the articulated skeleton's pose. These three matrices are calculated by linearizing the associated non-linear mappings at the current 3D skeleton and camera poses. Let us further denote $\mathbf{x}$ as the vector of rigid camera-to-body motion. The first order Taylor expansion about the current camera and body 3D poses set the linear relationship between the 2D motion and the 3D motion as follows

$$\mathbf{y} = \mathbf{P}\,\mathbf{D}\,\mathbf{x} + \mathbf{P}\,\mathbf{J}\,\boldsymbol{\Omega}. \tag{4}$$

The details of mathematical development to obtain equation 4 are described in the supplementary material. This equation, which was used in previous works [34, 40], considers the approximate observation model as being accurate. Such hypothesis can cause either a failure case of reconstructing ground-truth 3D shape or at best a non-accurate estimate if the noise level is not too high and the equality constraint still can explain a kinematic skeleton's motion. Let us assume $\boldsymbol{\epsilon}$ as the sum of the Taylor's expansion remainder, the noise in skeleton modeling and the noise in 2D joints detection. If we consider $realObs$ as the vector of real 2D measurement, then the above equation can be rewritten as

$$\mathbf{z} = \mathbf{P}\,\mathbf{D}\,\mathbf{x} + \mathbf{P}\,\mathbf{J}\,\boldsymbol{\Omega} + \boldsymbol{\epsilon}. \tag{5}$$

In this work, we consider minimizing the ℓ_2-norm of the noise $\boldsymbol{\epsilon}$ instead of imposing the equality constraint 4 since in real application $\mathbf{y}$ is not available.

4. Robust 3D human motion from monocular 2D landmark motion

Given a skeleton of known and noisy 2D motion landmarks $\mathbf{z} \in \mathbb{R}^{2l}$ with sparse differential kinematics, the 3D motion reconstruction problem can be stated as follows

$$\left(\hat{\mathbf{x}}, \hat{\boldsymbol{\Omega}}\right) \in \underset{\left(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\Omega}}\right)}{\arg\min}\, \mathcal{L}\left(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\Omega}}\right) \tag{6}$$

$$\text{s.t. } \mathcal{L}\left(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\Omega}}\right) = \left\{\alpha \left\|\tilde{\boldsymbol{\Omega}}\right\|_1 + \frac{1}{2}\left\|\mathbf{z} - \mathbf{P}\,\mathbf{D}\tilde{\mathbf{x}} - \mathbf{P}\,\mathbf{J}\,\tilde{\boldsymbol{\Omega}}\right\|_2^2\right\} \tag{7}$$

Where α is real positive number weighting how sparse the differential kinematic will be. When α is large it enforces differential kinematic sparsity. The lasso optimization stated in equation 6 is a convex problem. It is non-differentiable every time one of the differential kinematic angle does not undergo any motion.

5. A Posteriori Recovery Condition of Ground-Truth Motion

Let us consider $(\mathbf{x}^\star, \boldsymbol{\Omega}^\star)$ as the ground-truth solution of the noise free equality problem

$$\mathbf{z} = \mathbf{P}\,\mathbf{D}\mathbf{x}^\star + \mathbf{P}\,\mathbf{J}\,\boldsymbol{\Omega}^\star. \tag{8}$$

The ground-truth support of $\boldsymbol{\Omega}^\star$ is then denoted $\mathcal{V}^\star$. Let us denote $\boldsymbol{\Omega}^\star_{\mathcal{V}^\star} \in \mathbf{R}^{|\mathcal{V}^\star|}$ the vector built from considering only non-elements of $\boldsymbol{\Omega}^\star$. Let us denote $\mathbf{J}_{\mathcal{V}^\star} \in \mathbf{R}^{2l\times|\mathcal{V}^\star|}$ the matrix constructed from $\mathbf{J}$ by keeping only the columns that correspond to non-zero elements of $\boldsymbol{\Omega}^\star$. The following theorem provides a sufficient tool of checklist to verify the ground-truth recovery of the noise-free motion 8 by solving problem 6.

Theorem 1 *The solution $\boldsymbol{\Omega}^\star$ of equation 8 is the unique minimizer of problem 6 if the following conditions are verified:*

1. $(\mathbf{P}\mathbf{J}_{\mathcal{V}^\star})$ *is full rank of dimension* $2\times l$.
2. $\left|[\mathbf{P}\,\mathbf{J}]_k^{\mathrm{T}}\left(\mathbf{P}\mathbf{J}_{\mathcal{V}^\star}\right)^{+\mathrm{T}} \mathit{sign}\left(\boldsymbol{\Omega}^\star\right)\right| < 1$, *for all columns* $[\mathbf{P}\,\mathbf{J}]_k$ *not in* $(\mathbf{P}\mathbf{J}_{\mathcal{V}^\star})$.
3. $\mathit{sign}\left(\boldsymbol{\Omega}^\star_{\mathcal{V}^\star}\right) = \mathit{sign}\left(\boldsymbol{\Omega}^\star_{\mathcal{V}^\star} - \alpha\left(\mathbf{P}\mathbf{J}_{\mathcal{V}^\star}{}^{\mathrm{T}}\mathbf{P}\mathbf{J}_{\mathcal{V}^\star}\right)^{-1}\mathit{sign}\left(\boldsymbol{\Omega}^\star_{\mathcal{V}^\star}\right)\right)$ *for all* α *such that* $0 < \alpha < \bar{\alpha}$.
4. $\left(\mathbf{P}\mathbf{J}_{\mathcal{V}^\star}\right)^{+\mathrm{T}} \mathit{sign}\left(\boldsymbol{\Omega}^\star\right) \in \ker\left(\left(\mathbf{P}\,\mathbf{D}\right)^{\mathrm{T}}\right)$.

Please refer the supplementary material for the proof of this theorem.

6. Experimental Results

6.1. Implementation and Experimental Setup

In this work, we use the SMPL model as a kinematic skeleton model whose root is at the center of the hip [32]. The SMPL model is a parameteric human model with shape parameters and pose parameters. It has 23 pivot joints which provides a set of 69 angular kinematic degrees of freedom. The camera to skeleton's root orientation is represented by 3-rotations that are part of the $\boldsymbol{\Omega}$ vector. The total angular degrees of freedom is $n = 72$. $\mathbf{x}$ is a 3-real vector that represents the translation from camera to skeleton's root. The SMPL model has in addition 10 parameters that cast different human shapes and forms. In this work, we do not estimate these parameters and we use the ground-truth values provided in the dataset for both the proposed and compared methods. The proposed formulation is implemented using PYTHON3.7 on MACBOOKPRO running at 2.3 GHZ an INTEL CORE I9 processor. Problem 6 is solved using proximal gradient descent methods [12, 39]. The initial pose that is needed as first reference is computed using a gold standard method [37]. This method is applied on the first 3 frames of every sequence. These frames are discarded in the evaluation process. The weighting parameter α is experimentally set to 0.2. We use 3 publicly available dataset: SURREAL [52], HUMAN3.6M [25] and PANOPTIC [27]. Using SURREAL, we evaluate on all video sequences from the test set which contains 12528 clips of 100 frames each. This dataset contains synthetic human 3D poses with noise-free 2D landmarks. To provide addiontal tests on the robustness to noise in landmar detection, we augment the 2D landmarks of this dataset with a Gaussian noise of zero mean and 2 pixels in standard deviation. With HUMAN3.6M, the evaluation is processed on all actions for subject 9 and 11 as was initially used in [25]. The remaining subjects are reserved for training by the deep-learning methods. We use Protocol 1 which uses all the point of views and Protocol 2 which uses only the frontal point of view [5, 41, 51]. Finally, the pose 1 video scenarios from PANOPTIC are used as third set of frames for validation. We use an fps of 30 for sampling the videos of the three datasets. We show that this frame rate is high enough to keep descent model noise in the linearized equation 4 and guaranty a valid assumption of sparse articulated motion from frame-to-frame.

6.2. Evaluation

The proposed approach is compared to 4 state-of-the-art methods that consist in *[28]*, *[49]*, *[42]* and *[34]*. *[28]* infers 3D human from an image sequence by using a spatio-temporal deep-encoder. *[49]* produces 3D human reconstructions using an attention tempral Convolutional Neural Network. *[42]* uses a ConvNet to regress the 75 degrees of freedom of SMPL pose and 10 parameters of SMPL shape. In our study we compare the result with this method by using ground-truth SMPL shapes from the dataset. *[34]* uses an ℓ_1-norm minimization of the angular degrees of freedom with equality constraint as in equation 4. This method is used to show the relevance of the robust formulation of the proposed approach. Our quantitative study uses the Mean Per-Joint Position Error (MPJPE) in [mm] with the center hip as basis root. We also report the reconstruction error [19], which uses Procrustes Analysis to rigidly align the reconstruction with ground truth and then compute the MPJPE. Reconstruction error relates only the body posture to the ground truth without including the camera-to-body estimate. Tables 1 and 2 report the MPJPE and reconstruction errors for the compared and proposed methods. Table 2-left reports MPJPE and reconstruction errors considering noise-free 2D landmarks positions from SURREAL. Table 2-right reports MPJPE when adding noise in 2D landmarks positions. The added noise is a centered Gaussian

	MPJPE(P1)	MPJPE(P2)	Reconst. Error
[28]	87.7	79.3	58.6
[49]	88.5	52.1	
[42]	77.7	85.4	74.1
[34]	81.8	75.9	70.3
Ours	68.3	70.1	65.9

	MPJPE	Reconst. Error
[28]	58.4	43.7
[49]	66.5	54.8
[42]	56.9	53.1
[34]	26.8	19.8
Ours	23.1	14.9

Table 1. Top: Evaluation of the proposed method on HUMAN3.6M, following Protocol 1 (P1) and Protocol 2 (P2). Reconstruction error is reported as an average for both protocols. Bottom: Evaluation of the proposed method on PANOPTIC.

	MPJPE	Reconst. Error
[28]	65.7	48.7
[49]	71.2	57.8
[42]	34.6	22.3
[34]	28.9	23.14
Ours	25.1	19.4

Noise std	1	2	3	4
[28]	66.7	71.1	74.3	78.9
[49]	72.3	73.2	77.4	82.3
[42]	35.4	36.8	39.7	43.4
[34]	29.8	32.5	38.6	44.3
Ours	25.7	27.1	30.1	33.4

Table 2. Evaluation of the proposed method on SURREAL. Top: MPJPE and reconstruction errors zero noise in 2D landmarks. Bottom: MPJPE under different standard deviation of Gaussion centered noise in 2D landmarks.

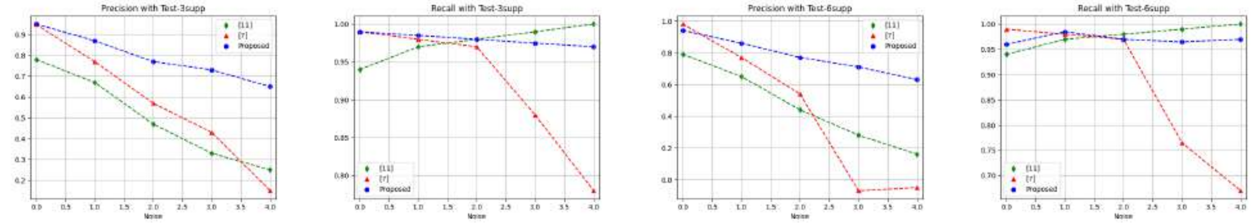

Figure 2. Precision and recall curves using synthetic sparse differential kinematics. Results of protocols Test-3supp and Test-6supp are reported.

with different standard deviations ranging from 1 to 4 pixels. The proposed method shows global good averages in MPJPE and reconstruction errors. It particularly handles well the challenging postures like Eat and Sit-Down in HUMAN3.6M. Figures 4 and 3 display qualitative results on challenging sequences from SURREAL dataset. As can be observed, the arm motion with upper movement in figure 3 and backward movement in figure 4 are better reconstructed with the proposed lasso formulation. Finally, SURREALdataset is used to evaluate the sufficient conditions of ground-truth recovery from theorem 1. The test set from this data is modified according to two protocols: *(1)* Test-3supp regenerate the 3D motion by setting to zero 69 elements of $\mathbf{\Omega}$. *(2)* Test-6supp regenerate the 3D motion by setting to zero 66 elements of $\mathbf{\Omega}$. For every test we ran experiments with different standard deviation noise in the 2D detected landmarks. Figures 2 draws the precision-recall curves of support recoveries for both protocols. It appears that the proposed methods shows robustness to increasing standard deviation noise. At low amount of noise (less than 3 pixels in std), the proposed method and [34] shows similar rate of accuracy. At higher amount of noise, [34] shows weaker precision due to the equality constraint that is less satisfied. Test-3supp and Test-6supp show also that in noise free case [34] performs better to recover angular motion with fewer sparsity. Test-3supp has given 95% rate of success in the sufficient conditions of theorem 1 while Test-6supp has decreased this rate to 80%. This result shows how much restrictive are the sufficient conditions and reveal the fact that they fit better motion recovery with very low amount of moving joints. The method by *[28]*shows weak precision performance and good recall performance. *[49]* and *[42]* are not displayed because they showed performances similar to *[28]*. It appears that deep neural network based approaches do not perform well to recover zero differential kinematics.

6.3. Discussion

The resolution of problem 6 allows us to accurately and robustly reconstruct 3D human motion from 2D landmark motion. The minimization of the reprojection constraint is more realistic then solving the problem with an equality constraint as was proposed in [34]. The paramerter α tunes how sparse the angular rotations should be. More α is small and close to zero, more the solution of 6 converges to the same solution provided by an equality constraint [18]. Experimentally, the threshold value is $\alpha_0 = 10^{-5}$. In all our experiments we used $\alpha = 0.2$. We test the effect of occluding randomly one 2D landmark per-frame during the whole sequence of SURREAL dataset. The results are as follows in term of MPJPE: *[28]*: 70.4, *[49]*: 74.9, *[42]*: 38.5, *[34]*:33.2 and *Ours*: 27.6. To avoid failure cases we re-compute an initialzation pose every 300 frames in average for a 30 fps video as was done in [34].

7. Conlculsion

In this paper we have presented a robust approach to reconstruct 3D human motion from noisy 2D detected landmarks. The proposed method relies on a lasso formulation with an ℓ_1-norm on the joint motion and an ℓ_2-norm on the reprojection error of the 2D motion. The lasso-based formulation shows better performance than a formulation based on an equality constraint of the reprojection error. This paper also presented a posteriori sufficient condition verification on the exact support recovery using a lasso-based formulation. The experiments have shown 95% of success in the sufficient conditions when the joint motion has 3 degrees of freedom of non-zero differential kinematics.

References

[1] A. Agarwal and B. Triggs. Recovering 3D human pose from monocular images. IEEE *Transactions on Pattern Analysis and Machine Intelligence*, 28(1), January 2006. 1

[2] Antonio Agudo and Francesc Moreno-Noguer. Robust Spatio-Temporal Clustering and Reconstruction of Multiple Deformable Bodies. *IEEE transactions on pattern analysis and machine intelligence*, 41(4):971–984, Apr. 2019. 1

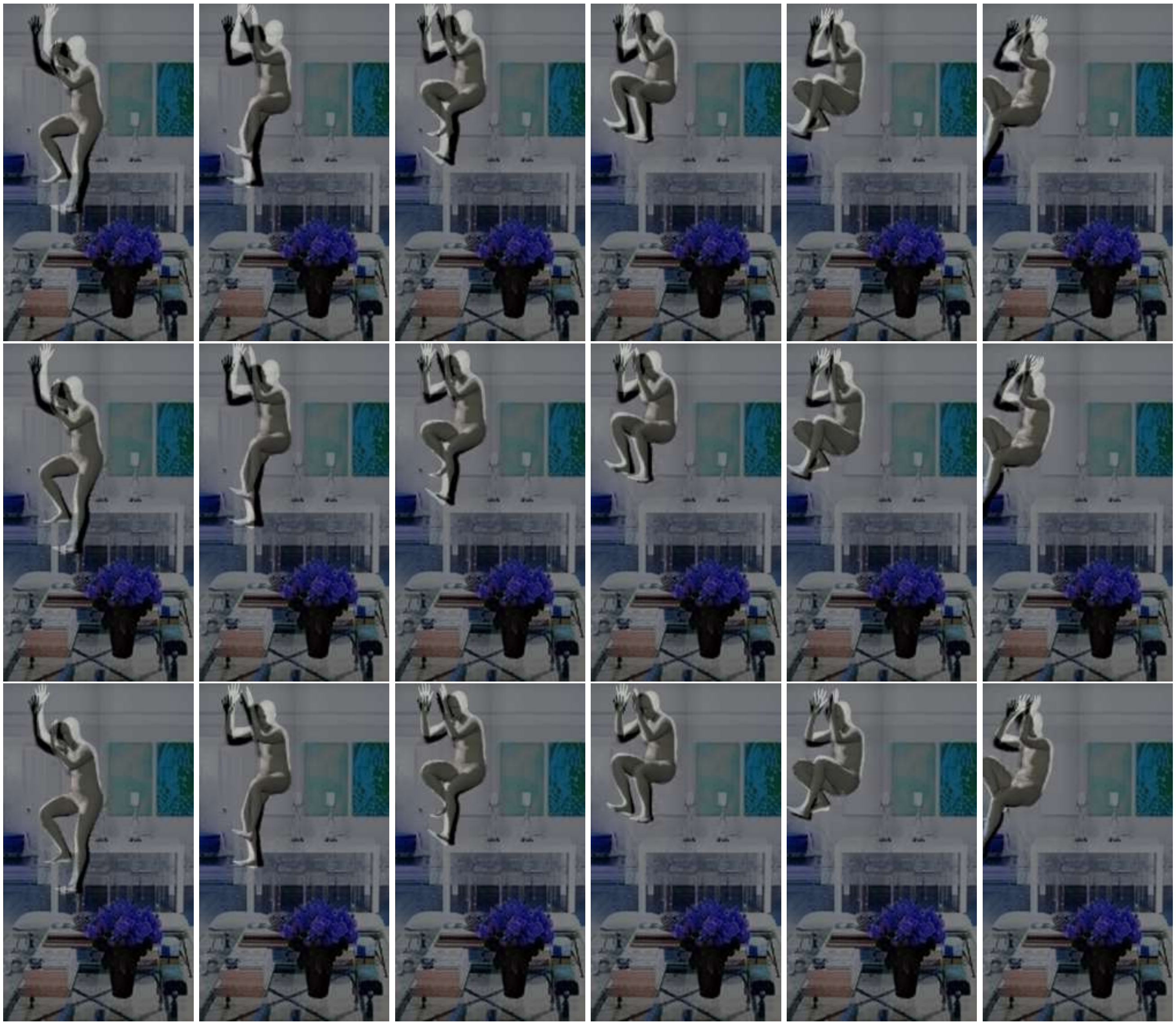

Figure 3. Qualitative reconstruction from SURREAL dataset. Session: run0/01_06_c0003. Frames from left to right: 40, 60, 70, 80, 90, 99. Top to bottom rows are reconstructions from *[28]*, *[34]*and *Ours*respectively.

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d Human Pose Estimation: New Benchmark and State of the Art Analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, June 2014. 1

[4] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting Temporal Context for 3D Human Pose Estimation in the Wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep It SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *ECCV 2016*, Lecture Notes in Computer Science, pages 561–578. Springer, Cham, Oct. 2016. 5

[6] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No.98CB36231)*, pages 8–15, June 1998. 1

[7] Z. Cao, T. Simon, S. E. Wei, and Y. Sheikh. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *2017 IEEE Conference on Computer Vision*

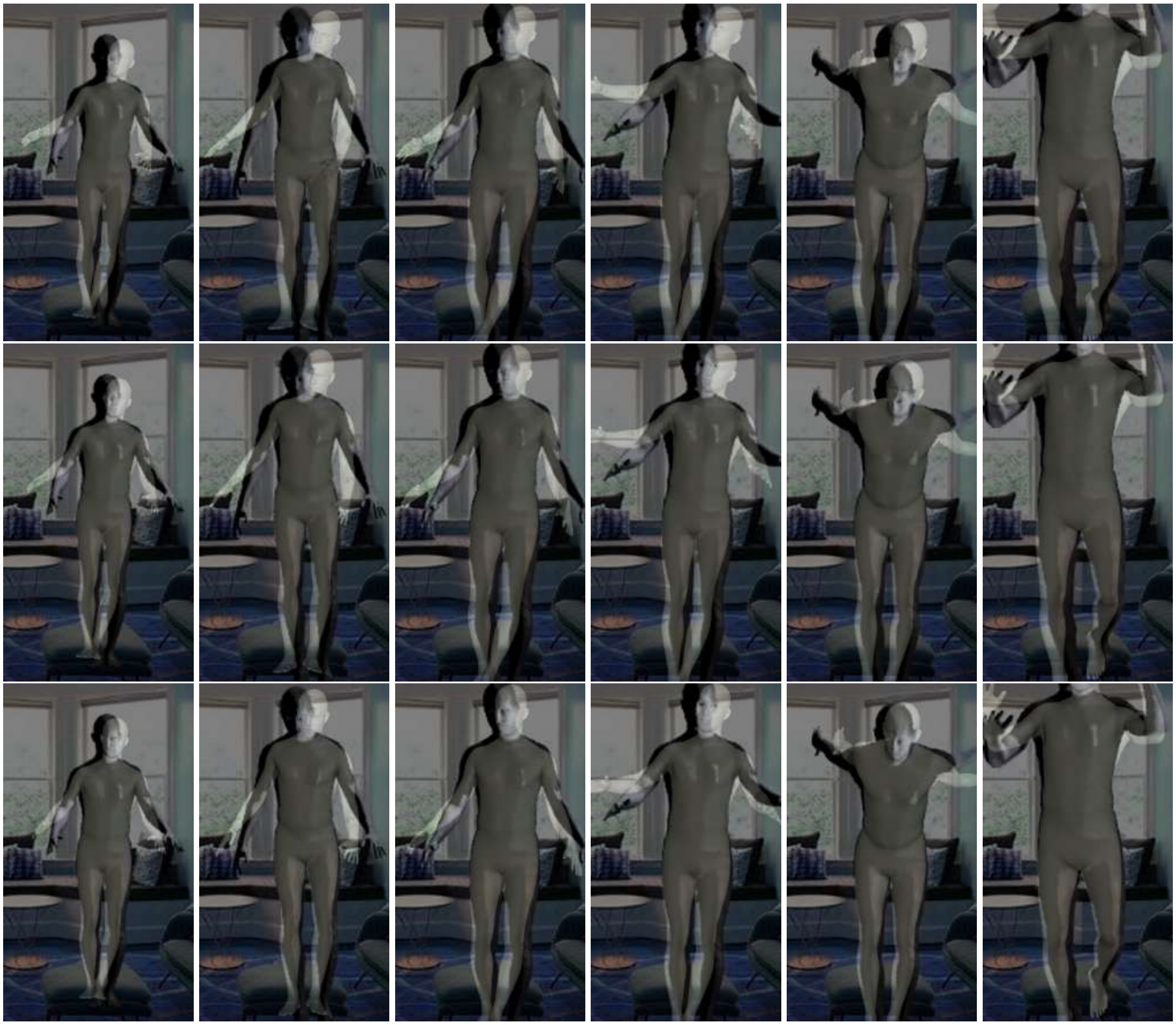

Figure 4. Qualitative reconstruction from SURREALdataset. Session: run0/106_04_c0001. Frames from left to right: 40, 60, 70, 80, 90, 99. Top to bottom rows are reconstructions from *[28]*, *[34]*and *Ours*respectively..

and Pattern Recognition (CVPR), pages 1302–1310, July 2017. 2

[8] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. SportsCap: Monocular 3D Human Motion Capture and Fine-Grained Understanding in Challenging Sports Videos. *International Journal of Computer Vision*, 129(10):2846–2864, Oct. 2021. 1

[9] Yen-Lin Chen and Jinxiang Chai. 3d Reconstruction of Human Motion and Skeleton from Uncalibrated Monocular Video. In *ACCV*, Lecture Notes in Computer Science, pages 71–82. Springer, Berlin, Heidelberg, Sept. 2009. 1

[10] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning To Estimate Robust 3D Human Mesh From In-the-Wild Crowded Scenes. pages 1475–1484, 2022. 1

[11] Vasileios Choutas, Lea Müller, Chun-Hao P. Huang, Siyu Tang, Dimitrios Tzionas, and Michael J. Black. Accurate 3D Body Shape Regression Using Metric and Semantic Attributes. pages 2718–2728, 2022. 2

[12] P. L. Combettes and Valérie R. Wajs. Signal Recovery by Proximal Forward-Backward Splitting. *Multiscale Model. Simul.*, 2005. 5

[13] Enric Corona, Gerard Pons-Moll, Guillem Alenyà,

and Francesc Moreno-Noguer. Learned Vertex Descent: A New Direction for 3D Human Model Fitting. Technical Report arXiv:2205.06254, arXiv, May 2022. arXiv:2205.06254 [cs] type: article. 3

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. SMPLicit: Topology-Aware Generative Model for Clothed People. pages 11875–11885, 2021. 1

[15] A. Del Bue, J. Xavier, L. Agapito, and M. Paladini. Bilinear Modeling via Augmented Lagrange Multipliers (BALM). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1496–1508, Aug. 2012. 1, 3

[16] A. Elgammal and Chan-Su Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–681–II–688 Vol.2, June 2004. 1

[17] Xiaochuan Fan, Kang Zheng, Youjie Zhou, and Song Wang. Pose Locality Constrained Representation for 3d Human Pose Reconstruction. In *Computer Vision ECCV 2014*, Lecture Notes in Computer Science, pages 174–188. Springer, Cham, Sept. 2014. 1

[18] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, June 2004. Conference Name: IEEE Transactions on Information Theory. 1, 6

[19] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 5

[20] Peng Guan, A. Weiss, A. O. Bălan, and M. J. Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388, Sept. 2009. 1

[21] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormahlen, and H. P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1823–1830, June 2010. 3

[22] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, pages 275–283, New York, NY, USA, Oct. 2019. Association for Computing Machinery. 1

[23] Yan Huang, Qiang Wu, JingSong Xu, Yi Zhong, and ZhaoXiang Zhang. Clothing Status Awareness for Long-Term Person Re-Identification. pages 11895–11904, 2021. 1

[24] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. ArtTrack: Articulated Multi-Person Tracking in the Wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1293–1301, July 2017. 2

[25] 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, July 2014. 5

[26] Yasamin Jafarian and Hyun Soo Park. Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos. pages 12753–12762, 2021. 1

[27] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3334–3342, Dec. 2015. 5

[28] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3D Human Dynamics From Video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5614–5623, 2019. 1, 3, 5, 6, 7, 8

[29] Isinsu Katircioglu, Bugra Tekin, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Learning Latent Representations of 3D Human Pose with Deep Neural Networks. *International Journal of Computer Vision*, pages 1–16, Jan. 2018. 2

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video Inference for Human Body Pose and Shape Estimation. pages 5253–5263, 2020. 3

[31] S. Leonardos, X. Zhou, and K. Daniilidis. Articulated motion estimation from a monocular image sequence using spherical tangent bundles. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 587–593, May 2016. 1

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, Oct. 2015. 1, 2, 4

[33] Zhichao Ma, Kan Li, and Yang Li. Self-supervised method for 3D human pose estimation with consistent shape and viewpoint factorization. *Applied Intelligence*, June 2022. 2

[34] Abed Malti. On the exact recovery conditions of 3D human motion from 2D landmark motion with sparse

articulated motion. *Computer Vision and Image Understanding*, 202:103072, Jan. 2021. 1, 3, 4, 5, 6, 7, 8

[35] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. *arXiv:1705.03098 [cs]*, May 2017. arXiv: 1705.03098. 2

[36] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D Human Pose Estimation In The Wild Using Improved CNN Supervision. *arXiv:1611.09813 [cs]*, Nov. 2016. arXiv: 1611.09813. 2

[37] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. VNect: Real-time 3D Human Pose Estimation with a Single RGB Camera. *ACM Trans. Graph.*, 36(4):44:1–44:14, July 2017. 1, 2, 3, 5

[38] F. Moreno-Noguer. 3D Human Pose Estimation from a Single Image via Distance Matrix Regression. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1561–1570, July 2017. 2

[39] Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa. Solving Structured Sparsity Regularization with Proximal Methods. In José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 418–433, Berlin, Heidelberg, 2010. Springer. 5

[40] H. S. Park and Y. Sheikh. 3D reconstruction of a smooth articulated trajectory from a monocular image sequence. In *2011 International Conference on Computer Vision*, pages 201–208, Nov. 2011. 1, 3, 4

[41] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[42] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to Estimate 3D Human Pose and Shape from a Single Color Image. *arXiv:1805.04092 [cs]*, May 2018. arXiv: 1805.04092. 2, 5, 6

[43] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3D Human Pose from 2D Image Landmarks. In *ECCV 2012*, Lecture Notes in Computer Science, pages 573–586. Springer, Berlin, Heidelberg, Oct. 2012. 1

[44] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR 2017 - IEEE Conference on Computer Vision & Pattern Recognition*, Honolulu, United States, June 2017. 2

[45] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2019. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 1, 2

[46] Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. HULC: 3D Human Motion Capture with Pose Manifold Sampling and Dense Contact Guidance. Technical Report arXiv:2205.05677, arXiv, May 2022. arXiv:2205.05677 [cs] type: article. 3

[47] L. Sigal and M. J. Black. HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, 2006. 1

[48] M. S. Subodh Raj and Sudhish N. George. 11/2 Regularized RPCA Technique for 3D Human Action Recovery. In *2020 IEEE 17th India Council International Conference (INDICON)*, pages 1–5, Dec. 2020. ISSN: 2325-9418. 1

[49] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei. Human Mesh Recovery From Monocular Images via a Skeleton-Disentangled Representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5349–5358, 2019. 3, 5, 6

[50] 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. 1

[51] D. Tome, C. Russell, and L. Agapito. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698, July 2017. 2, 5

[52] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning From Synthetic Humans. pages 109–117, 2017. 5

[53] B. Wandt, H. Ackermann, and B. Rosenhahn. 3d Reconstruction of Human Motion from Monocular Image Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1505–1516, Aug. 2016. 1, 3

[54] Bastian Wandt and Bodo Rosenhahn. RepNet: Weakly Supervised Training of an Adversarial Reprojection

Network for 3D Human Pose Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7782–7791, 2019. 1, 2

[55] Chunyu Wang, Yizhou Wang, Zhouchen Lin, and Alan L. Yuille. Robust 3D Human Pose Estimation from Single Images or Video Sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1227–1241, May 2019. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 3

[56] C. Wang, Y. Wang, Z. Lin, A. L. Yuille, and W. Gao. Robust Estimation of 3d Human Poses from a Single Image. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2369–2376, June 2014. 1

[57] Yuanhao Wu and Chenxing Wang. Parallel-branch network for 3D human pose and shape estimation in video. *Computer Animation and Virtual Worlds*, n/a(n/a):e2078. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.2078. 3

[58] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3D Human Pose Estimation With Spatial and Temporal Transformers. pages 11656–11665, 2021. 3

[59] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep Kinematic Pose Regression. *ECCV Workshop on Geometry Meets Deep Learning*, Sept. 2016. arXiv: 1609.05317. 2

[60] X. Zhou, M. Zhu, S. Leonardos, K. G. Derpanis, and K. Daniilidis. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4966–4975, June 2016. 1

[61] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):901–914, Apr. 2019. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence. 2

[62] Y. Zhu, M. Cox, and S. Lucey. 3d motion reconstruction for real-world camera motion. In *CVPR 2011*, pages 1–8, June 2011. 1, 3