A Kinematic Chain Space for Monocular Motion Capture

Bastian Wandt, Hanno Ackermann, and Bodo Rosenhahn

Institut für Informationsverarbeitung, Leibniz Universität Hannover

Abstract. This paper deals with motion capture of kinematic chains (e.g. human skeletons) from monocular image sequences taken by uncalibrated cameras. We present a method based on projecting an observation onto a kinematic chain space (KCS). An optimization of the nuclear norm is proposed that implicitly enforces structural properties of the kinematic chain. Unlike other approaches our method is not relying on training data or previously determined constraints such as particular body lengths. The proposed algorithm is able to reconstruct scenes with little or no camera motion and previously unseen motions. It is not only applicable to human skeletons but also to other kinematic chains for instance animals or industrial robots. We achieve state-of-the-art results on different benchmark databases and real world scenes.

1 Introduction

Monocular human motion capture is an important and large part of recent research. Its applications range from surveillance, animation, robotics to medical research. While there exists a large number of commercial motion capture systems, monocular 3D reconstruction of human motion plays an important role where complex hardware arrangements are not feasible or too costly.

Recent approaches to the non-rigid structure from motion problem [1–4] achieve good results for laboratory settings. They are designed to work with tracked 2D points from arbitrary 3D point clouds. To resolve the duality of camera and point motion they require sufficient camera motion in the observed sequence. On the other hand, in many applications (e.g. human motion capture, animal tracking or robotics) properties of the tracked objects are known. Exploiting known structural properties for non-rigid structure from motion problems is rarely considered e.g. by using example based modeling as in [5] or constancy of bone lengths in [6]. Recently, linear subspace training approaches have been proposed [6–11]. They can efficiently represent human motion, even for 3D reconstruction from single images. However, they require extensive training on known motions which restricts them to reconstructions of the same motion category. Further, training based approaches cannot recover individual subtleties in the motion (e.g. limping instead of walking) sufficiently well.

This paper closes the gap between non-rigid structure from motion and subspace-based human modeling. Similar to other approaches which depend on

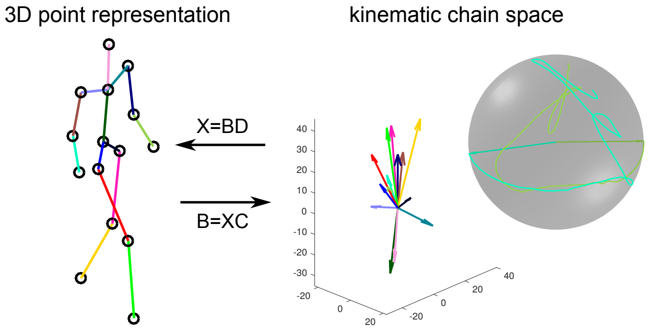

Fig. 1. Mapping from a 3D point representation to the kinematic chain space. The vectors in the KCS equal to directional vectors in the 3D point representation. The sphere shows the trajectories of left and right lower arm in KCS. Since both bones have the same length their trajectories lie on the same sphere.

the work of Bregler et al. [12], we decompose an observation matrix in three matrices corresponding to camera motion, transformation and basis shapes. Unlike other works that find a transformation which enforces properties of the camera matrices, we develop an algorithm that optimizes the transformation with respect to structural properties of the observed object. This reduces the amount of camera motion necessary for a good reconstruction. We experimentally found that even sequences without camera motion can be reconstructed. Unlike other works in the field of human modeling we propose to first project the observations in a *kinematic chain space (KCS)* before optimizing a reprojection error with respect to our kinematic model. Fig. 1 shows the mapping between the KCS and the representation based on 2D or 3D feature points. It is done by multiplication with matrices which implicitly encode a kinematic chain (cf. Sec. 3.1). This representation enables us to derive a nuclear norm optimization problem which can be solved efficiently. Imposing a low rank constraint on a Gram matrix has shown to improve 3D reconstructions [3]. However, the method of [3] is only based on constraining the camera motion. Therefore, it requires sufficient camera motion. The KCS allows to use a geometric constraint which is based on the topology of the underlying kinematic chain. Thus, the required amount of camera motion is much lower.

We evaluate our method on different standard databases (CMU MoCap [13], KTH [14], HumanEva [15], Human3.6M [16]) as well as on our own databases qualitatively and quantitatively. The proposed algorithm achieves state-of-the-art results and can handle problems like motion transfers and unseen motion. Due to the noise robustness of our method we can apply a CNN-based joint labeling algorithm [17, 18] for RGB images as input data which allows us to directly reconstruct human poses from unlabeled videos. Although this method is developed for human motion capture it is applicable to other kinematic chains such as animals or industrial robots as shown in the experiments in Sec. 4.3.

Summarizing, our contributions are:

- We propose a method for 3D reconstruction of kinematic chains from monocular image sequences.
- An objective function based on structural properties of kinematic chains is derived that not only imposes a low-rank assumption on the shape basis but also has a physical interpretation.
- We propose using a nuclear norm optimization in a *kinematic chain space.*
- In contrast to other works our method is not limited to previously learned motion patterns and does not use strong anthropometric constraints such a-priorly determined bone lengths.

2 Related Work

The idea of decomposing a set of 2D points tracked over a sequence into matrices whose entries are identified with the parameters of shape and motion was first proposed by Tomasi and Kanade [19]. A generalization of this algorithm to deforming shapes was proposed by Bregler et al. [12]. They assume that the observation matrix can be factorized into two matrices representing camera motion and multiple basis shapes. After an initial decomposition is found by singular value decomposition (SVD) of the observation matrix they compute a transformation matrix by enforcing camera constraints. Xiao et al. [20] showed that the basis shapes of [12] are ambiguous. They solved this ambiguity by employing basis constraints on them. As shown by Akther et al. [1] these basis constraints are still not sufficient to resolve the ambiguity. Therefore, they proposed to use an object independent trajectory basis. Torresani et al. [21–23] proposed to use different priors on the transformation matrix such as additional rank constraints and Gaussian priors. Gotardo and Martinez [24] built on the idea of [1] by applying the DCT representation to enforce a smooth 3D shape trajectory. Parallel to this work they proposed a solution that uses the kernel trick to also model nonlinear deformations [25] which cannot be represented by a linear combination of basis shapes. Hamsici et al. [2] also assume a smooth shape trajectory and apply the kernel trick to learn a mapping between the 3D shape and the 2D input data. Park et al. [26] introduced activity-independent spatial and temporal constraints. Inspired by [1] and [26] Valmadre et al. [27] proposed a dynamic programming approach combined with temporal filtering. Dai et al. [3] minimize the trace norm of the transformation matrix to impose a sparsity constraint. Different to [3] Lee et al. [28] define additional constraints on motion parameters to avoid the sparsity constraint. Since all these methods assume to work for arbitrary non-rigid 3D objects, none of them utilizes knowledge about the underlying kinematic structure. Rehan et al. [4] were the first to define a temporary rigidity of reconstructed structures by factorizing a small number of consecutive frames. Thereby, they can reconstruct kinematic chains if the object does not deform much. Due to their sliding window assumption, the method is even more restricted to scenes with sufficient camera motion.

Several works consider the special case of 3D reconstruction of human motion from monocular images. A common approach is to previously learn base poses of the same motion category. These are then linearly combined for the estimation of 3D poses. To avoid implausible poses, most authors utilize properties of human skeletons to constrain a reprojection error based optimization problem. However, anthropometric priors such as the sum of squared bone lengths [7], known limb proportions [8], known skeleton parameters [5], previously trained joint angle constraints [9] or strong physical constraints [29] all suffer from the fact that parameters have to be known a-priorly. Zhou et al. [10] propose a convex relaxation of the commonly used reprojection error formulation to avoid the alternating optimization of camera and object pose. While many approaches try to reconstruct human poses from a single image [30–35] using anthropometric priors, such constraints have rarely been used for 3D reconstruction from image sequences. Wandt et al. [6] constrain the temporal change of bone length without using a predefined skeleton. Zhou et al. [36] combined a deep neural network that estimates 2D landmarks with 3D reconstruction of the human pose. A different approach is to include sensors as additional information source [37–39]. Other works use a trained mesh model for instance SMPL [40] and project it to the image plane [41, 42]. The restriction to a trained subset of possible human motions is the major downside of these approaches.

In this paper we combine NR-SfM and human pose modeling without requiring previously learned motions. By using a representation that implicitly models the kinematic chain of a human skeleton our algorithm is capable to reconstruct unknown motion from labeled image sequences.

3 Estimating Camera and Shape

The i-th joint of a kinematic chain is defined by a vector $\boldsymbol{x}_i \in \mathbb{R}^3$ containing the x,y,z-coordinates of the location of this joint. By concatenating j joint vectors we build a matrix representing the pose $\boldsymbol{X}$ of the kinematic chain

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_j). \tag{1}$$

The pose $\boldsymbol{X}_k$ in frame k can be projected into the image plane by

$$\boldsymbol{X}'_k = \boldsymbol{P}_k \boldsymbol{X}_k, \tag{2}$$

where $\boldsymbol{P}_k$ is the projection matrix corresponding to a weak perspective camera. For a sequence of f frames, the pose matrices are stacked such that $\boldsymbol{W} = (\boldsymbol{X}'_1, \boldsymbol{X}'_2, \ldots, \boldsymbol{X}'_f)^T$ and $\hat{\boldsymbol{X}} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_f)^T$. This implies

$$\boldsymbol{W} = \boldsymbol{P}\hat{\boldsymbol{X}}, \tag{3}$$

where $\boldsymbol{P}$ is a block diagonal matrix containing the camera matrices $\boldsymbol{P}_{1,\ldots,f}$ for the corresponding frame. After an initial camera estimation we subtract a matrix $\boldsymbol{X}_0$ from the measurement matrix by

$$\hat{\boldsymbol{W}} = \boldsymbol{W} - \boldsymbol{P}\hat{\boldsymbol{X}}_0, \tag{4}$$

where $\hat{\boldsymbol{X}}_0$ is obtained by stacking $\boldsymbol{X}_0$ multiple times to obtain the same size as $\boldsymbol{W}$. Here, we take $\boldsymbol{X}_0$ to be a mean pose. We will provide experimental evidence that the algorithm proposed in the following is insensitive w.r.t. the choice of $\boldsymbol{X}_0$ as long as it represents a reasonable configuration of the kinematic chain. In all the experiments dealing with kinematic chains of humans, we take $\boldsymbol{X}_0$ to be the average of all poses in the CMU data set.

Following the approach of Bregler et al. [12] we decompose $\hat{\boldsymbol{W}}$ by Singular Value Decomposition to obtain a rank-$3K$ pose basis $\boldsymbol{Q} \in \mathbb{R}^{3K \times j}$. While [12] and similar works then optimize a transformation matrix with respect to orthogonality constraints of camera matrices, we optimize the transformation matrix with respect to constraints based on a physical interpretation of the underlying structure. With $\boldsymbol{A}$ as transformation matrix for the pose basis we may write

$$\boldsymbol{W} = \boldsymbol{P}(\hat{\boldsymbol{X}}_0 + \boldsymbol{A}\boldsymbol{Q}). \tag{5}$$

In the following sections we will present how poses can be projected into the kinematic chain space (Sec. 3.1) and how we derive an optimization problem from it (Sec. 3.2). Combined with the camera estimation (Sec. 3.3) an alternating algorithm is presented in Sec. 3.4.

3.1 Kinematic Chain Space

To define a bone $\boldsymbol{b}_k$, a vector between the r-th and t-th joint is computed by

$$\boldsymbol{b}_k = \boldsymbol{p}_r - \boldsymbol{p}_t = \boldsymbol{X}\boldsymbol{c}, \tag{6}$$

where

$$\boldsymbol{c} = (0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^T, \tag{7}$$

with 1 at position r and -1 at position t. The vector $\boldsymbol{b}_k$ has the same direction and length as the corresponding bone. Similarly to Eq. (1), a matrix $\boldsymbol{B} \in \mathbb{R}^{3 \times b}$ can be defined containing all b bones

$$\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_b). \tag{8}$$

The matrix $\boldsymbol{B}$ is calculated by

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{C}, \tag{9}$$

where $\boldsymbol{C} \in \mathbb{R}^{j \times b}$ is built by concatenating multiple vectors $\boldsymbol{c}$. Analogously to $\boldsymbol{C}$, a matrix $\boldsymbol{D} \in \mathbb{R}^{b \times j}$ can be defined that maps $\boldsymbol{B}$ back to $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{D}. \tag{10}$$

$\boldsymbol{D}$ is constructed similar to $\boldsymbol{C}$. Each column adds vectors in $\boldsymbol{B}$ to reconstruct the corresponding point coordinates. Note that $\boldsymbol{C}$ and $\boldsymbol{D}$ are a direct result of the underlying kinematic chain. Therefore, the matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ perform the mapping from point representation into the *kinematic chain space* and vice versa.

3.2 Trace Norm Constraint

One of the main properties of human skeletons is the fact that bone lengths do not change over time.

Let

$$\boldsymbol{\Psi} = \boldsymbol{B}^T\boldsymbol{B} = \begin{pmatrix} l_1^2 & \cdot & \cdot & \cdot \\ \cdot & l_2^2 & \cdot & \cdot \\ \cdot & \cdot & \ddots & \cdot \\ \cdot & \cdot & \cdot & l_b^2 \end{pmatrix}. \tag{11}$$

be a matrix with the squared bone lengths on its diagonal. From $\boldsymbol{B} \in \mathbb{R}^{3\times b}$ follows $rank(\boldsymbol{B}) = 3$. Thus, $\boldsymbol{\Psi}$ has rank 3. Note that if $\boldsymbol{\Psi}$ is computed for every frame we can define a stronger constraint on $\boldsymbol{\Psi}$. Namely, as bone lengths do not change for the same person the diagonal of $\boldsymbol{\Psi}$ remains constant.

Proposition 1. *The nuclear norm of* $\boldsymbol{B}$ *is invariant for any bone configuration of the same person.*

Proof. The trace of $\boldsymbol{\Psi}$ equals the sum of squared bone lengths (Eq. (11))

$$trace(\boldsymbol{\Psi}) = \sum_{i=1}^{b} l_i^2. \tag{12}$$

From the assumption that bone lengths of humans are invariant during a captured image sequence the trace of $\boldsymbol{\Psi}$ is constant. The same argument holds for $trace(\sqrt{\boldsymbol{\Psi}})$. Therefore, we have

$$\|\boldsymbol{B}\|_* = trace(\sqrt{\boldsymbol{\Psi}}) = const. \tag{13}$$

Since this constancy constraint is non-convex we will relax it to derive an easy to solve optimization problem. Using Eq. (9) we project Eq. (5) into the KCS which gives

$$\boldsymbol{WC} = \boldsymbol{P}(\hat{\boldsymbol{X}}_0\boldsymbol{C} + \boldsymbol{AQC}) \tag{14}$$

The unknown is the transformation matrix $\boldsymbol{A}$. For better readability we define $\boldsymbol{B}_0 = \boldsymbol{X}_0\boldsymbol{C}$ and $\boldsymbol{S} = \boldsymbol{QC}$.

Proposition 2. *The nuclear norm of the transformation matrix* $\boldsymbol{A}$ *for each frame has to be greater than some scalar c, which is constant for each frame.*

Proof. Let $\boldsymbol{B} = \boldsymbol{B}_1 + \boldsymbol{B}_0$ be a decomposition of $\boldsymbol{B}$ into the initial bone configuration $\boldsymbol{B}_0$ and a difference to the observed pose $\boldsymbol{B}_1$. It follows that

$$\|\boldsymbol{B}\|_* = \|\boldsymbol{B}_1 + \boldsymbol{B}_0\|_* = c_1, \tag{15}$$

where c_1 is a constant. The triangle inequality for matrix norms gives

$$\|\boldsymbol{B}_1\|_* + \|\boldsymbol{B}_0\|_* \geq \|\boldsymbol{B}_1 + \boldsymbol{B}_0\|_* = c_1. \tag{16}$$

Since $\boldsymbol{B}_0$ is known, it follows

$$\|\boldsymbol{B}_1\|_* \geq c_1 - \|\boldsymbol{B}_0\|_* = c, \tag{17}$$

where c is constant. $\boldsymbol{B}_1$ can be represented in the shape basis $\boldsymbol{S}$ (cf. Sec. 3) by multiplying it with the transformation matrix $\boldsymbol{A}$

$$\boldsymbol{B}_1 = \boldsymbol{A}\boldsymbol{S}. \tag{18}$$

Since the shape base matrix $\boldsymbol{S}$ is a unitary matrix the nuclear norm of $\boldsymbol{B}_1$ equals

$$\|\boldsymbol{B}_1\|_* = \|\boldsymbol{A}\|_*. \tag{19}$$

By Eq. (17) follows that

$$\|\boldsymbol{A}\|_* \geq c. \tag{20}$$

Proposition 2 also holds for a sequence of frames. Let $\hat{\boldsymbol{A}}$ be a matrix built by stacking $\boldsymbol{A}$ for each frame and $\hat{\boldsymbol{B}}_0$ be defined similarly, we relax Eq. (20) and obtain the final formulation for our optimization problem

$$\begin{aligned} \min_{\hat{\boldsymbol{A}}} \quad & \|\hat{\boldsymbol{A}}\|_* \\ \text{s.t.} \quad & \|\boldsymbol{W}\boldsymbol{C} - \boldsymbol{P}(\hat{\boldsymbol{A}}\boldsymbol{S} + \hat{\boldsymbol{B}}_0)\|_F = 0. \end{aligned} \tag{21}$$

Eq. (21) does not only define a low rank assumption on the transformation matrix. By the derivation above, we showed that the nuclear norm is reasonable because it has a concise physical interpretation. More intuitively, the minimization of the nuclear norm will give solutions close to a mean configuration $\boldsymbol{B}_0$ of the bones in terms of rotation of the bones. The constraint in Eq. (21) which represents the reprojection error prevents the optimization from converging to the trivial solution $\|A\|_* = 0$. This allows for a reconstruction of arbitrary poses and skeletons.

Moreover, Eq. (21) is a well studied problem which can be efficiently solved by common optimization methods such as Singular Value Thresholding (SVT) [43].

3.3 Camera

The objective function in Eq. (21) can also be optimized for the camera matrix $\boldsymbol{P}$. Since $\boldsymbol{P}$ is a block diagonal matrix, Eq. (21) can be solved block-wise for each frame. With $\boldsymbol{X}_i'$ and $\boldsymbol{P}_i$ corresponding to the observation and camera at frame i the optimization problem can be written as

$$\min_{\boldsymbol{P}_i} \|\boldsymbol{X}_i'\boldsymbol{C} - \boldsymbol{P}_i(\boldsymbol{A}\boldsymbol{S} + \boldsymbol{B}_0)\|_F. \tag{22}$$

Considering the entries in

$$\boldsymbol{P}_i = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \end{pmatrix} \tag{23}$$

we can enforce a weak perspective camera by the constraints

$$p_{11}^2 + p_{12}^2 + p_{13}^2 - (p_{21}^2 + p_{22}^2 + p_{23}^2) = 0 \tag{24}$$

and

$$p_{11}p_{21} + p_{12}p_{22} + p_{13}p_{23} = 0. \tag{25}$$

3.4 Algorithm

In the previous sections we derived an optimization problem that can be solved for the camera matrix $\boldsymbol{P}$ and transformation matrix $\boldsymbol{A}$ respectively. As both are unknown we propose algorithm 1 which alternatingly solves for both matrices. Initialization is done by setting all entries in the transformation matrix $\boldsymbol{A}$ to zero. Additionally, an initial bone configuration $\boldsymbol{B}_0$ is required. It has to roughly model a human skeleton but does not need to be the mean of the sequence.

Algorithm 1 Factorization algorithm for kinematic chains

```
% Input:
$\boldsymbol{B}_0 \leftarrow$ initial bone configuration
$\boldsymbol{C} \leftarrow$ kinematic chain matrix
$\boldsymbol{W} \leftarrow$ observation
$f \leftarrow$ number of frames
$\boldsymbol{A} \leftarrow \mathbf{0}$

while no convergence do
    for $t = 1 \rightarrow f$ do
        optimize $\|\boldsymbol{X}_t\boldsymbol{C} - \boldsymbol{P}_t(\boldsymbol{A}\boldsymbol{S} + \boldsymbol{B}_0)\|_F$
        insert $\boldsymbol{P}_t$ in $\boldsymbol{P}$
    end for
    perform SVT on
        min $\|\hat{\boldsymbol{A}}\|_*$ s.t. $\|\boldsymbol{W}\boldsymbol{C} - \boldsymbol{P}(\hat{\boldsymbol{A}}\boldsymbol{S} + \hat{\boldsymbol{B}}_0)\|_F = 0$
end while

% Output:
$\boldsymbol{P}$: camera matrices
$(\hat{\boldsymbol{A}}\boldsymbol{S} + \hat{\boldsymbol{B}}_0)\boldsymbol{D}$: 3D poses
```

4 Experiments

For the evaluation of our algorithm different benchmark data sets (CMU MoCap [13], HumanEva [15], KTH [14], Human3.6M [16]) were used. As measure for the quality of the 3D reconstructions we calculate the *Mean Per Joint Position Error (MPJPE)* [44] which is defined by

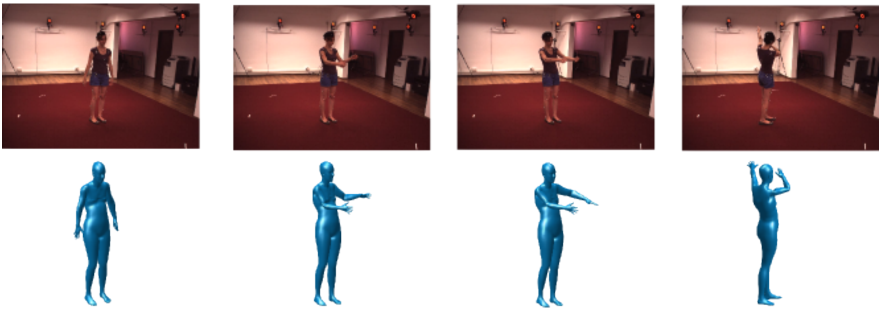

Fig. 2. Reconstruction of the highly articulated *directions* sequence from the Human3.6M data set subject 1.

$$e = \frac{1}{j} \sum_{i=1}^{j} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|, \tag{26}$$

where $\boldsymbol{x}_i$ and $\hat{\boldsymbol{x}}_i$ correspond to the ground truth and estimated positions of the i-th joint respectively. By rigidly aligning the 3D reconstruction to the ground truth we obtain the *3D positioning error (3DPE)* as introduced by [45]. To compare sequences of different lengths the mean of the 3DPE over all frames is used. In the following it is referred to as *3D error*.

Additional to this quantitative evaluation we perform reconstructions of different kinematic chains in Sec. 4.3 and on unlabeled image sequences in Sec. 4.4. All animated meshes in this section are created using SMPL [40]. The SMPL model is fitted to the reconstructed skeleton and is used solely for visualization.

4.1 Evaluation on Benchmark Databases

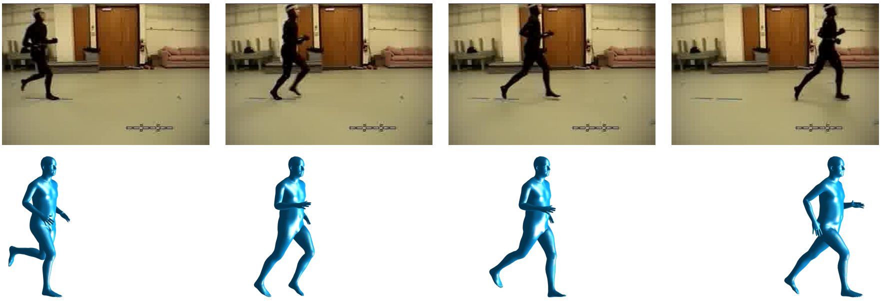

Fig. 3. Reconstruction of a running motion from the CMU database subject 35/17.

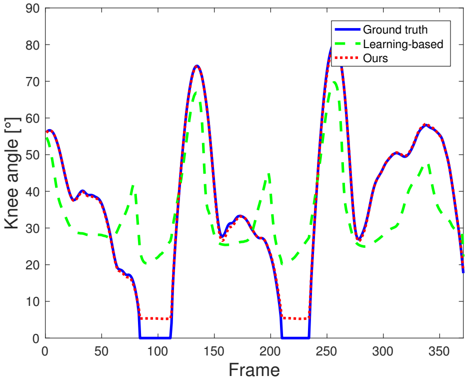

Fig. 4. Knee angle of reconstructions of a limping motion. The learning-based method [6] struggles to reconstruct minor differences from the motion patterns used for training whereas our learning-free approach recovers the knee angle in more detail.

To qualitatively show the drawbacks of learning-based approaches we reconstructed a sequence of a limping person. We use the method of [6] trained on walking patterns to reconstruct the 3D scene. Although the motions are very similar, the algorithm of [6] is not able to reconstruct the subtle motions of the limping leg. Fig. 4 shows the knee angle of the respective leg. The learning-based method reconstructs a periodic walking motion and cannot recover the unknown asymmetric motion which makes it unusable for gait analysis applications. The proposed algorithm is able to recover the motion in more detail.

We compare our method with the unsupervised works [1, 2] and the learning-based approach of [6]. The codes of [1] and [2] are freely available. Although there are slightly newer works, these two approaches show the inherent problem of these unsupervised methods (as also shown in [4]). We are not aware of any works that are able to reconstruct scenes with very limited or no camera motion without a model of the underlying structure. Rehan et al. [4] assume a local rigidity that allows for defining a kinematic chain model. This reduced the amount of necessary camera motion to 2 degrees per frame. However, due to their assumption that the observed object is approximately rigid in a small time window they are limited to a constantly moving camera.

For each sequence we created 20 random camera paths with little or no camera motion and compared our 3D reconstruction results with the other methods. Table 1 shows the 3D error in *mm* for different sequences and data sets. For the entry *walk35* we calculated the mean overall 3D errors of all 23 walking sequences from subject 35 in the CMU database. The columns *jump* and *limp* show the 3D error of a single jumping and limping sequence. *KTH* means the football sequence of the KTH data set [14] and *HE* the walking sequence of the HumanEva data set [15]. The last four columns are average errors over all subjects performing the respective motions of the Human3.6M data set [16]. Note that the highly articulated motions from Human3.6M data set vary a lot in the same category and therefore are hard to learn by approaches like [6]. All these sequences are captured with little or no camera motion. The unsupervised methods of [1] and [2] require more camera motion and completely fail in these scenarios. The learning-based approach of [6] reconstructs plausible poses for all sequences. They even achieve a better result for the walking motions. However, motions with larger variations between persons and sequences (e.g. jumping and limping) are harder to reconstruct from the learned pose basis. Although the results look like plausible human motions, they lack the ability to reconstruct subtle motion variations. In contrast, the proposed method is able to reconstruct these variations and achieves a better result. Some of our reconstructions are shown in Figs. 2 and 3 for sequences of the Human3.6M and CMU data set, respectively.

Table 1. 3D error in *mm* for different sequences and data sets. The column *walk35* shows the mean 3D error of all sequences containing walking motion from subject 35 in the CMU database. *jump* refers to the jumping motion of subject 13/11 of the CMU database and *limp* to the limping motion of subject 91/16. *KTH* means the football sequence of the KTH data set [14]. The column *HE* shows the 3D error for the HumanEva walking sequence [15]. The last four columns are average errors over all subjects performing the respective motions of the Human3.6M data set [16].

	walk35	jump	limp	KTH	HE	3.6M walk	3.6M dir.	3.6M pose	3.6M photo
[1]	228.68	210.14	99.37	108.91	106.92	86.76	130.43	121.33	145.44
[2]	264.75	186.70	112.92	114.03	102.99	**66.70**	121.40	120.56	136.30
[6]	**11.22**	45.49	64.46	68.88	58.62	71.54	110.36	135.87	124.52
Ours	18.94	**36.50**	**19.24**	**53.10**	**44.36**	74.44	**80.83**	**109.28**	**101.76**

4.2 Convergence

We alternatingly optimize the camera matrices (Eq. (21)) and transformation matrix (Eq. (22)). Since convergence of the algorithm cannot be guaranteed we show it by experiment. Fig. 5 shows the convergence of the reprojection error in pixel for a sequence from the CMU MoCap database. However, the reprojection error only shows the convergence of the proposed algorithm but cannot prove that the 3D reconstructions will improve every iteration. We additionally estimated the convergence of the 3D error in Fig. 5. In most cases our algorithm converges to a good minimum in less than 3 iterations. Further iterations do not improve the visual quality and only deform the 3D reconstruction less than $1mm$. The 3D error remains constant during camera estimation which causes the *steps* in the error plot.

Fig. 6 shows the computation time over the number of frames for three different sequences. The computation time mostly depends on the number frames and less on the observed motion. We use unoptimized Matlab code on a desktop PC for all computations.

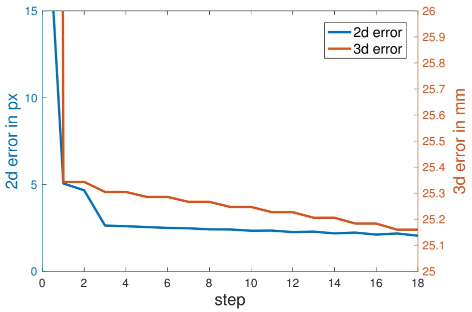

Fig. 5. Reprojection error and 3D error with respect to number of iterations for subject35/sequence1 from the CMU MoCap data set. Even steps refer to camera estimation while odd steps correspond to shape estimation.

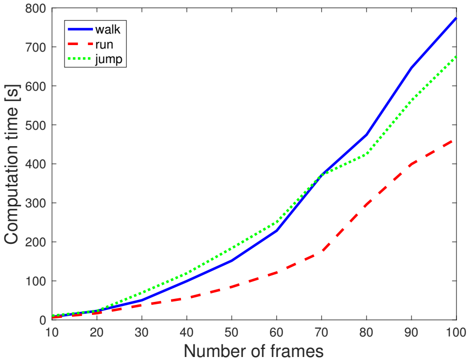

Fig. 6. Computation time for walking, running and jumping sequences of the CMU data set using unoptimized Matlab code. It mostly depends on the number of frames and less on the observed motion.

4.3 Other Kinematic chains

Although our method was developed for the reconstruction of human motion, it generalizes to all kinematic chains that do not include translational joints. In this section we show reconstructions of other kinematic chains such as people holding objects, animals and industrial robots.

In situations where people hold objects with both hands the kinematic chain of the body can be extended by another rigid connection between the two hands. Fig. 7 shows the reconstruction of the sword fighting sequence of the CMU data set. By simply adding another column to the kinematic chain space matrix $\boldsymbol{C}$ (cf. Sec. 3.1) the distance between the two hands is enforced to remain constant. The exact distance does not need to be known, however.

Fig. 8 shows a robot used for precision milling and the reconstructed 3D model as overlay. The proposed method is able to correctly reconstruct the robots motion. In Fig. 9 we reconstructed a more complex motion of a horse during show

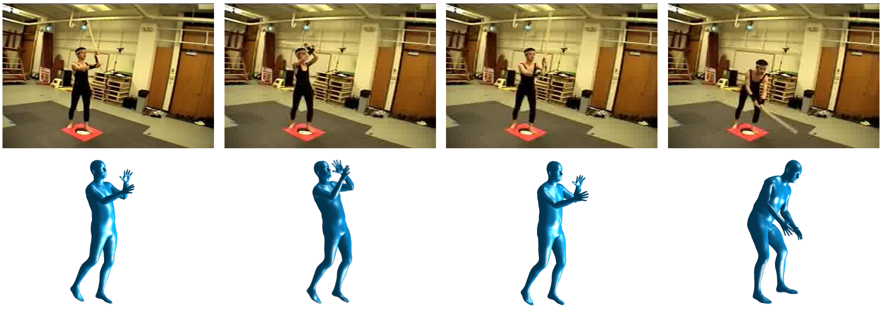

Fig. 7. Reconstruction of the sword play sequence of the CMU database. The kinematic chain is extended such that the hands are rigidly connected.

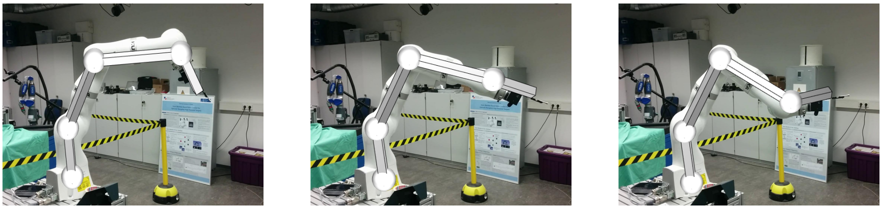

Fig. 8. Reconstruction of a sequence of an industrial robot moving along a path. The reconstruction is shown as an augmented overlay over the images.

jumping. We used a simplified model of the bone structure of a horse. Also in reality the shoulder joint is not completely rigid. Despite these limitations the algorithm achieves plausible results.

4.4 Image Sequences

The proposed method is designed to reconstruct a 3D object from labeled feature points. In the former sections this was done by setting and tracking them semi-interactively. In this section we will show that our method is also able to use the noisy output of a human joint detector. We use *deeperCut* [17, 18] to estimate the joints in the outdoor run and jump sequence from [46]. Fig. 10 shows the joints estimated by *deeperCut* and our 3D reconstruction. As can be seen in Fig. 10 we achieve plausible 3D reconstructions even with automatically labeled noisy input data.

5 Conclusion

We developed a method for the 3D reconstruction of kinematic chains from monocular image sequences. By projecting into the kinematic chain space a constraint is derived that is based on the assumption that bone lengths are constant.

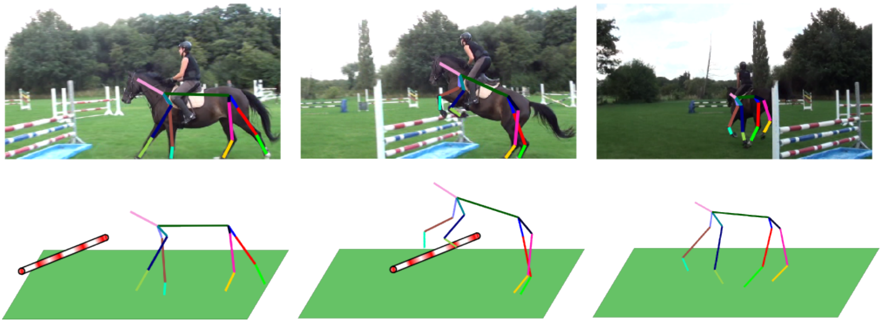

Fig. 9. Reconstruction of a horse riding sequence. Although we use a very rough model for the skeleton of the horse we obtain plausible reconstructions. The complete reconstruction including more views can be seen in the supplemental video.

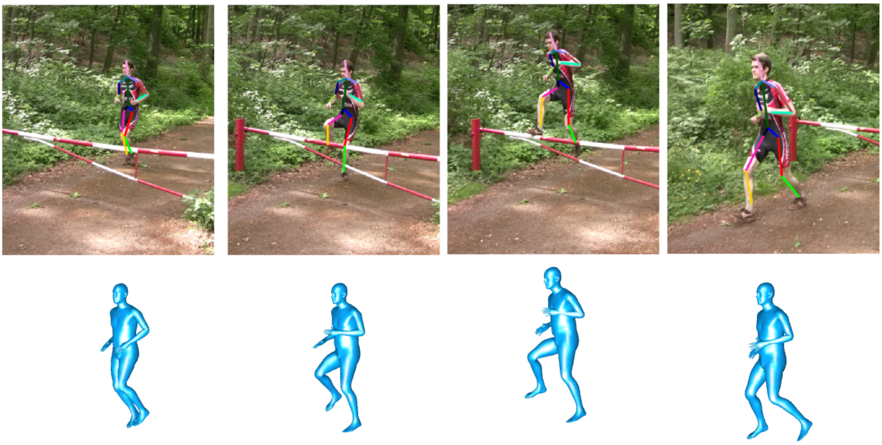

Fig. 10. Reconstruction of a running and jumping sequence from [46] automatically labeled by *deeperCut* [17, 18].

This results in the formulation of an easy to solve nuclear norm optimization problem. It allows for reconstruction of scenes with little camera motion where other non-rigid structure from motion methods fail. Our method does not rely on previous training or predefined body measures such as known limb lengths. The proposed algorithm generalizes to the reconstruction of other kinematic chains and achieves state-of-the-art results on benchmark data sets.

References

1. Akhter, I., Sheikh, Y., Khan, S., Kanade, T.: Trajectory space: A dual representation for nonrigid structure from motion. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(7) (7 2011) 1442–1456
2. Hamsici, O., Gotardo, P., Martinez, A.: Learning spatially-smooth mappings in non-rigid structure from motion. In: European Conference on Computer Vision (ECCV). (2011)
3. Dai, Y., Li, H.: A simple prior-free method for non-rigid structure-from-motion factorization. In: Conference on Computer Vision and Pattern Recognition (CVPR). CVPR '12, Washington, DC, USA, IEEE Computer Society (2012) 2018–2025
4. Rehan, A., Zaheer, A., Akhter, I., Saeed, A., Mahmood, B., Usmani, M., Khan, S.: Nrsfm using local rigidity. In: Proceedings Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, IEEE (March 2014) 69–74
5. Chen, Y.L., Chai, J.: 3d reconstruction of human motion and skeleton from uncalibrated monocular video. In Zha, H., Taniguchi, R.I., Maybank, S.J., eds.: Asian Conference on Computer Vision (ACCV). Volume 5994 of Lecture Notes in Computer Science., Springer (2009) 71–82
6. Wandt, B., Ackermann, H., Rosenhahn, B.: 3d reconstruction of human motion from monocular image sequences. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8) (2016) 1505–1516
7. Ramakrishna, V., Kanade, T., Sheikh, Y.A.: Reconstructing 3d human pose from 2d image landmarks. In: European Conference on Computer Vision (ECCV). (October 2012)
8. Wang, C., Wang, Y., Lin, Z., Yuille, A., Gao, W.: Robust estimation of 3d human poses from a single image. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
9. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2015). (June 2015) 1446–1455
10. Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3d shape estimation from 2d landmarks: A convex relaxation approach. In: CVPR, IEEE Computer Society (2015) 4447–4455
11. Wandt, B., Ackermann, H., Rosenhahn, B.: 3d human motion capture from monocular image sequences. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. (June 2015)
12. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2000) 690–696
13. CMU: Human motion capture database (2014)
14. Kazemi, V., Burenius, M., Azizpour, H., Sullivan, J.: Multi-view body part recognition with random forests. In: British Machine Vision Conference (BMVC). (2013)
15. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision **87**(1-2) (2010) 4–27
16. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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) (jul 2014) 1325–1339

17. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
18. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: European Conference on Computer Vision (ECCV). (2016)
19. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision **9** (1992) 137–154
20. Xiao, J., Chai, J., Kanade, T.: A closed-form solution to non-rigid shape and motion recovery. In: European Conference on Computer Vision (ECCV). (May 2004)
21. Torresani, L., Hertzmann, A., Bregler, C.: Learning non-rigid 3d shape from 2d motion. In Thrun, S., Saul, L.K., Schölkopf, B., eds.: Neural Information Processing Systems (NIPS), MIT Press (2003)
22. Torresani, L., Hertzmann, A., Bregler., C.: Nonrigid structure-from-motion: Estimating shape and motion with hierarchical priors. IEEE Transactions Pattern Analysis and Machine Intelligence (2008)
23. Torresani, L., Yang, D.B., Alexander, E.J., Bregler, C.: Tracking and modeling non-rigid objects with rank constraints. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2001) 493–500
24. Gotardo, P., Martinez, A.: Non-rigid structure from motion with complementary rank-3 spaces. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2011)
25. Gotardo, P., Martinez, A.: Kernel non-rigid structure from motion. In: International Conference on Computer Vision (ICCV), IEEE (2011)
26. Park, H.S., Sheikh, Y.: 3d reconstruction of a smooth articulated trajectory from a monocular image sequence. In Metaxas, D.N., Quan, L., Sanfeliu, A., Gool, L.J.V., eds.: ICCV, IEEE Computer Society (2011) 201–208
27. Valmadre, J., Zhu, Y., Sridharan, S., Lucey, S.: Efficient articulated trajectory reconstruction using dynamic programming and filters. In Fitzgibbon, A.W., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: ECCV (1). Volume 7572 of Lecture Notes in Computer Science., Springer (2012) 72–85
28. Lee, M., Cho, J., Choi, C.H., Oh, S.: Procrustean normal distribution for non-rigid structure from motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 1280–1287
29. Zell, P., Wandt, B., Rosenhahn, B.: Joint 3d human motion capture and physical analysis from monocular videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (July 2017)
30. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings IEEE International Conference on Computer Vision (ICCV), Piscataway, NJ, USA, IEEE (October 2017)
31. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (July 2017)
32. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
33. Rogez, G., Weinzaepfel, P., Schmid, C.: LCR-Net: Localization-Classification-Regression for Human Pose. In: CVPR 2017 - IEEE Conference on Computer

Vision & Pattern Recognition, Honolulu, United States, IEEE (July 2017) 1216–1224

34. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: A weakly-supervised approach. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
35. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Computer Vision and Pattern Regognition (CVPR). (2018)
36. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
37. von Marcard, T., Pons-Moll, G., Rosenhahn, B.: Human pose estimation from video and imus. Transactions on Pattern Analysis and Machine Intelligence **38**(8) (January 2016) 1533–1547
38. von Marcard, T., Rosenhahn, B., Black, M., Pons-Moll, G.: Sparse inertial poser: Automatic 3d human pose estimation from sparse imus. Computer Graphics Forum 36(2), Proceedings of the 38th Annual Conference of the European Association for Computer Graphics (Eurographics) (2017)
39. von Marcard, T., Henschel, R., Black, M.J., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: European Conference on Computer Vision. (2018)
40. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Trans. Graphics (Proc. SIGGRAPH Asia) **34**(6) (October 2015) 248:1–248:16
41. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Springer International Publishing (October 2016)
42. Alldieck, T., Kassubeck, M., Wandt, B., Rosenhahn, B., Magnor, M.: Optical flow-based 3d human motion estimation from monocular video. In: German Conference on Pattern Recognition (GCPR). (September 2017)
43. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization **20**(4) (March 2010) 1956–1982
44. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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) (2014) 1325–1339
45. Simo-Serra, E., Ramisa, A., Aleny, G., Torras, C., Moreno-Noguer, F.: Single image 3d human pose estimation from noisy observations. In: Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2012) 2673–2680
46. Hasler, N., Rosenhahn, B., Thormählen, T., Wand, M., Seidel, H.P.: Markerless motion capture with unsynchronized moving cameras. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)