EXPRESS PAPER   **Open Access**

Learning 3D joint constraints from vision-based motion capture datasets

Pramod Murthy[1,2*†], Hammad T. Butt[1,3†], Sandesh Hiremath[2], Alireza Khoshhal[1] and Didier Stricker[1,2]

Abstract

Realistic estimation and synthesis of articulated human motion must satisfy anatomical constraints on joint angles. A data-driven approach is used to learn human joint limits from 3D motion capture datasets. We represent joint constraints with a new formulation (s_1, s_2, τ) using swing-twist representation in exponential maps form. Our parameterization is applied on Human3.6M dataset to create the lookup-map for each joint. These maps enable us to generate 'synthetic' datasets in entire joint rotation space of a given joint. A set of neural network discriminators is then trained with synthetic datasets to learn valid/invalid joint rotations. The discriminators achieve accuracy of $[94.4 - 99.4\%]$ for different joints. We validate precision-accuracy trade-off of discriminators and qualitatively evaluate classified poses with an interactive tool. The learned discriminators can be used as 'priors' for human pose estimation and motion synthesis.

Keywords: Human pose estimation, Motion capture, Joint constraint

1 Introduction

3D human pose estimation from monocular images using kinematic models [1–3] has produced state of art *robust* results compared to direct 2D-to-3D regression models [4, 5]. An enormous collection of human motion capture (mocap) datasets are now publicly available to train deep neural networks. Recently, [6] has even employed a synthetic human motion dataset for better training of deep neural networks. But, the synthesis of *realistic* human motion for long run is an open problem [7].

Often, a *'penalty term'* is used for exceeding joint limits, but these joint limits are worked out from a limited dataset [8, 9]. To the best of our knowledge, no recent attempt has been reported to obtain more realistic and statistically valid joint limits from publicly available mocap datasets.

The artifacts and outliers in the mocap data can be pruned using joint constraints inferred directly from respective datasets. The imposition of accurate joint constraints as *priors* is valuable for realistic 3D human pose

estimation and motion synthesis. A data-driven learning of joint constraints from mocap datasets is a challenging task. The presence of outliers and voids due to sparse data makes it imperative to choose a representation of joint space, which is robust to such artifacts. Moreover, the instances of class representing *invalid joint rotations* are missing in existing motion capture datasets. One-class learning on these dataset, using positive-unlabelled (PU) algorithms has associated challenges. Instead, we *synthesize* both valid and invalid samples for balanced learning.

The main contributions of our paper are (i) set of neural network discriminators to classify valid and invalid rotations for different joints (ii) a new formulation of swing-twist(s_1, s_2, τ) representation to represent joint rotations (iii) discretized look-up map based on above parameterization to synthesize labels in a given joint space.

The overview of the current work for the left shoulder joint is depicted in the Fig. 1. Our three-parameter representation is unambiguous and free of anomalies, which makes it easier to learn a joint rotation manifold. Thus, neural network discriminators trained on our representation yield good accuracy and qualitative precision with very few parameters.
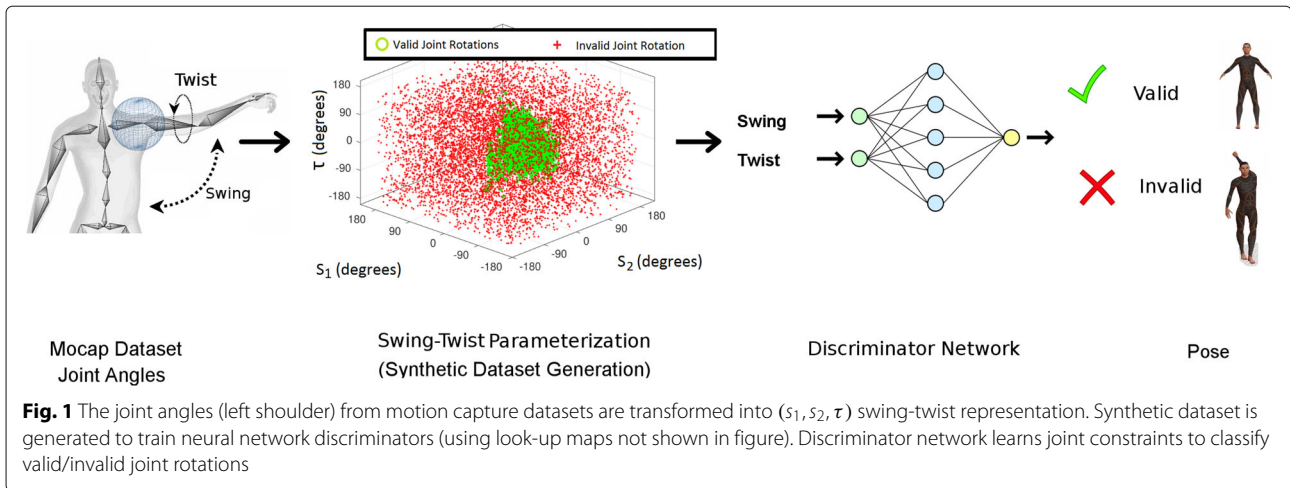
*Correspondence: pramod.murthy@dfki.de
†Pramod Murthy and Hammad T. Butt contributed equally to this work.
[1]Department of Augmented Vision, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Trippstadter Straße 122, 67663 Kaiserslautern, Germany
[2]TU Kaiserslautern, Kaiserslautern, Germany
Full list of author information is available at the end of the article

Fig. 1 The joint angles (left shoulder) from motion capture datasets are transformed into (s_1, s_2, τ) swing-twist representation. Synthetic dataset is generated to train neural network discriminators (using look-up maps not shown in figure). Discriminator network learns joint constraints to classify valid/invalid joint rotations

2 Related works

The joint constraints are dependent on the choice of representation for joint rotations. A straightforward approach to apply joint constraints in terms of Euler angles is to assume minimum and maximum limits on each angle (box-model) [10]. However, this approach is too crude to capture realistic anatomical joint limits. Euler angles representation also leads to discontiguous cluster of valid poses in joint space and often causes the problem of Gimbal lock [11]. Due to these limitations, the joint limits based on box model are not optimal choice. The swing-twist representation in [11–14] of joint rotation is devised to efficiently classify validity of a joint rotation. However, interdependence of swing and twist (intra-joint dependence) has not been satisfactorily addressed in earlier works.

In a data-driven approach, Herda et al. [15] has proposed a parameterization in terms of an implicit surface on joint trajectory data in quaternion space. The authors extend their method to deal with inter-joint dependencies[1], but it is computationally expensive and is not intuitive to interpret the joint limits. Further, the imposed limits are directly on captured data, which might lead to unrealistic limits especially in the presence of voids and sparse mocap data. Also, only the vector part $(\sin \theta)$ of quaternion is used in [15] to represent joint limits. It clearly ignores information in scalar part $(\cos \theta)$ leading to ambiguity in rotation angle of joint. The quaternion also has dual representation $(+q$ and $-q)$ for a given joint rotation. A preprocessing step is often necessary over motion trajectory to mitigate the effects of duality.

Recently, Akhter et al. [9] learned pose-conditioned joint angle limits of trained gymnasts and athletes from a more controlled mocap dataset. Their work encodes only swing of a body segment in a discretized binary occupancy map. The occupancy map is defined using local spherical coordinates of the joint on a unit sphere. A table-look-up function for validity of pose is conditioned for every child

joint based on its immediate parent in a kinematic chain while ignoring the intra-joint dependency[1].

Aforementioned limitations of previous works form the motivation of our research. We use an exponential map parameterization instead to represent joint rotations in present work. The exponential map representation avoid singularities as well as non-contiguous clusters [10]. It also avoids the duality problem in representation as described earlier.

We derive our joint constraints using more comprehensive Human 3.6 Million(H36M) dataset [16]. To overcome the limitation of intra-joint dependency, an additional parameter representing twist is learned. We encode lower and upper bounds of twist τ around a joint axis on occupancy map for a given swing (s_1, s_2).

3 Proposed methodology

The constraints of a 3D human joint rotation for routine activities are learned by using Human3.6M (H36M) dataset [16]. Our joint constraints parameterization is based on swing-twist formulation. Initially, we decompose the joint rotation (expressed in quaternion form) into swing and twist parts. Later, we express each of the swing and twist in exponential map and arrive at (s_1, s_2, τ).

In comparison, Akhter et al. [9] learn joint rotation constraints from a motion capture dataset based on extreme range of joint movements. Moreover, they ignore twist around axis as they parameterize joint rotation from local 3D Cartesian coordinates to the spherical coordinates(ϕ, θ) representing swing of body segment. In current work, we compare our joint limits (swing part) derived from a range of normal activities performed by human subjects (in H36M) with extreme range limits provided by [9].

3.1 Swing-twist decomposition

The decomposition of joint rotation into orthogonal components, i.e., swing and twist is achieved by defining the

twist w.r.t an axis of intrinsic (rotated) frame of reference. For example, if y-axis of reference frame is aligned with child limb, the swing-twist decomposition of a rotation in terms of quaternions is expressed in the following way: The quaternion $q := (q_s, q_x, q_y, q_z)$ can be decomposed as $q_{twist} * q_{swing}$, where $q_{twist} := (c_y, 0, s_y, 0)$, $q_{swing} := (c_{zx}, s_x, 0, s_z)$. Using the norm constraint of unit quaternion and following [13] we obtain,

$$
\begin{aligned}
c_y &= \left(\frac{q_s}{\sqrt{q_s^2 + q_y^2}} \right), s_y = \left(\frac{q_y}{\sqrt{q_s^2 + q_y^2}} \right), c_{zx} = \left(\frac{(q_s^2 + q_y^2)}{\sqrt{q_s^2 + q_y^2}} \right) \\
s_x &= \left(\frac{(q_s \cdot q_x - q_y \cdot q_z)}{\sqrt{q_s^2 + q_y^2}} \right), s_z = \left(\frac{(q_s \cdot q_z + q_x \cdot q_y)}{\sqrt{q_s^2 + q_y^2}} \right)
\end{aligned} \tag{1}
$$

Then, we introduce $\eta_{twist} := (0, \tau, 0)$ and $\eta_{swing} := (s_1, 0, s_2)$ as respective exponential map counter-parts, and get,

$$
\begin{aligned}
\tau &= 2 \cdot \arctan\left(\frac{s_y}{c_y} \right), \\
s_1 &= 2 \cdot \arctan\left(\frac{\sqrt{s_x^2 + s_z^2}}{c_{zx}} \right) \cdot \frac{s_x}{\sqrt{s_x^2 + s_z^2}}, \\
s_2 &= 2 \cdot \arctan\left(\frac{\sqrt{s_x^2 + s_z^2}}{c_{zx}} \right) \cdot \frac{s_z}{\sqrt{s_x^2 + s_z^2}}
\end{aligned} \tag{2}
$$

Using the Eq. (2), we get swing components (s_1, s_2) and the twist component τ of a joint rotation in exponential map. This representation is convenient as it (i) is unambiguous (no duality), (ii) avoids non-contiguous regions/jumps in joint trajectory, and (iii) requires no additional constraints like unit norm. Compared with swing-twist decomposition in quaternion form [13], it requires less parameters (3 parameters in η_{twist}, η_{swing} versus 5 in q_{twist} and q_{swing}). Consequently, it provides optimal joint representation for learning valid/invalid poses.

4   Implementation
The Euler angle representation of joint angles in H36M and [9] dataset are transformed into (s_1, s_2, τ) representation, as described in previous section. The resulting joint rotation space is continuous, contiguous and well-behaved for discriminative learning (Fig. 2).

It is found that considerable outliers are present in mocap dataset provided by Akhter et al. [9]. For instance, left and right knee flexion of all subjects performing various activities is shown in Fig. 3. It clearly shows an impractical knee extension (upward) of up to $-20°$ for certain motion sequences in dataset. Moreover, we believe that limited set of mocap data of [9] is not enough to learn inter-joint dependency, i.e., insufficient instances of swing

of the child limb exists conditioned upon a prior swing of parent limb.

The comparison of swing limits for right hip joint (right up leg) from both datasets in swing space (s_1, s_2) is illustrated in Fig. 2. The swing limits obtained using H36M although more restrictive but are true representative of human motion involved in routine activities. Also, these limits lie well within the extreme range of motion of [9]. Therefore, we restrict our learning only to H36M dataset.

4.1   Look-up maps for synthetic dataset
Let $\mathbb{M}$ be the set of all samples of H36M dataset for joint J, represented in (s_1, s_2, τ) space . We define a 2D grid (s^i, s^j) on swing subspace, where $s^i, s^j \in \{-180, \dots, 180\}$ and a discrete mapping as follows

$$
D_s(s_1, s_2) := \begin{cases} (s^i, s^j) & \text{iff } s_1 \in [s^i, s^{i+1}] \\ & \text{and } s_2 \in [s^j, s^{j+1}] \\ \text{undefined} & \text{otherwise} \end{cases} \tag{3}
$$

Now for each grid square (s^i, s^j), if at least one mocap sample $\in \mathbb{M}$ lies in it, we term it as valid swing bin. The discretized swing occupancy map is thus defined by

$$
O_s(s^i, s^j) := \begin{cases} 1 & \text{iff } \exists\, (s_1, s_2, \tau) \in \mathbb{M} \\ & \text{such that } D_s(s_1, s_2) = (s^i, s^j) \\ 0 & \text{otherwise} \end{cases} \tag{4}
$$

With each valid occupancy bin, we further associate minimum and maximum bounds of τ, based on all the instances of mocap data occuring in that bin

$$
L_t(s^i, s^j) := \begin{cases} [\tau_{min}, \tau_{max}] & \forall\, (s_1, s_2, \tau) \in \mathbb{M} \\ & \text{s.t.} O_s(s^i, s^j) = 1, \\ & \tau_{min} = \inf(\tau) \text{ and} \\ & \tau_{max} = \sup(\tau) \\ \text{undefined} & \text{otherwise} \end{cases} \tag{5}
$$

In order to obtain smooth twist angle bounds, we apply 3×3 Gaussian kernel(s) on (s_1, s_2, τ_{min}) and (s_1, s_2, τ_{max}) surfaces. Now given any (s_1, s_2, τ), we define a validity function in the following way

$$
\text{isValid}(s_1, s_2, \tau) = \begin{cases} 1 & \text{iff } O_s \circ D_s(s_1, s_2) = 1 \\ & \text{and } \tau \in [\tau_{min}, \tau_{max}], \\ & [\tau_{min}, \tau_{max}] = L_t \circ D_s(s_1, s_2) \\ 0 & \text{otherwise} \end{cases} \tag{6}
$$

We generate synthetic ground truth of labeled samples in (s_1, s_2, τ) space of a joint J by uniform random sampling on $s_1, s_2, \tau \in [-180, 180]$ and using isValid() function defined in Eq. 6.

4.2   Training discriminators
A large portion is occupied by invalid instances in complete joint rotation space (for any given joint). Therefore,
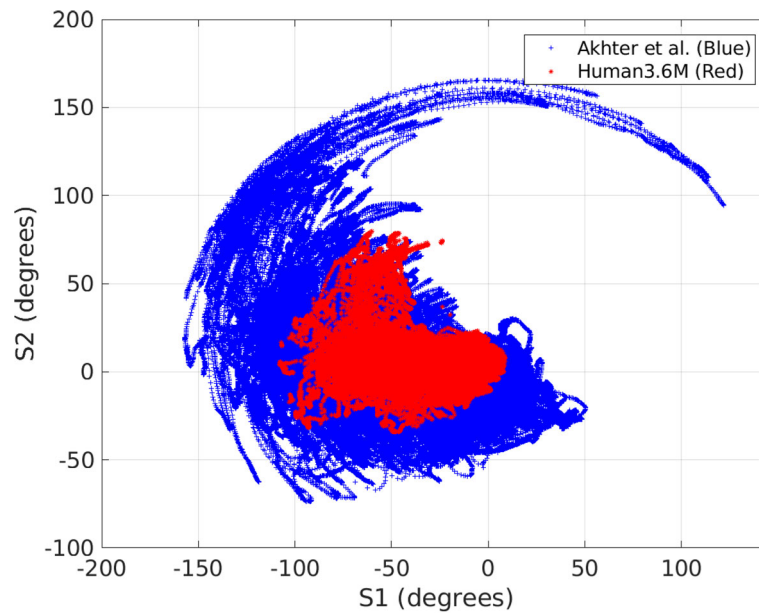
Fig. 2 The comparison of right hip joint's range of swing present in H36M [16] and Akhter et al. [9] dataset

the sets representing valid and invalid classes are imbalanced. We perform random sampling of valid and invalid instances to obtain 25%/75% sets respectively. The sampled sets are equally spaced 'labeled' training instances in entire joint space. An example of such sampled joint space for right hip joint (right up leg) is as shown in Fig. 4.

A balanced (50%/50%) subset is selected from these sampled valid/invalid instances using random sampling.

The subset is again divided into training, validation, and test samples in a ratio (70%, 15%, 15%). For comparison, we also tested over-sampling of valid class and under-sampling of invalid class to obtain balanced subset independently. However, the sampling proposed above leads to least variation of performance, when evaluation is performed on multiple test sets. The neural network discriminator are trained for each joint, using an equal
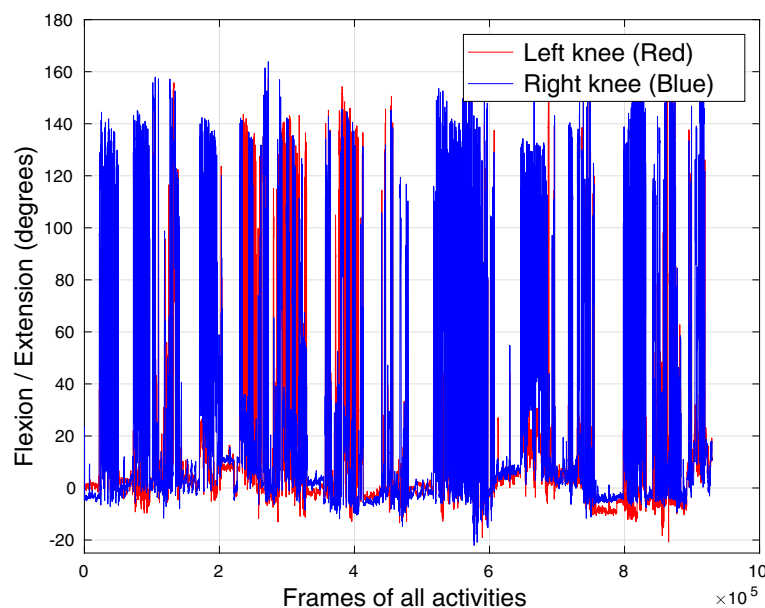
Fig. 3 The knee joint rotation anomalies in Akhter et al. dataset [9]. Flexion is positive and extension is negative knee joint angle
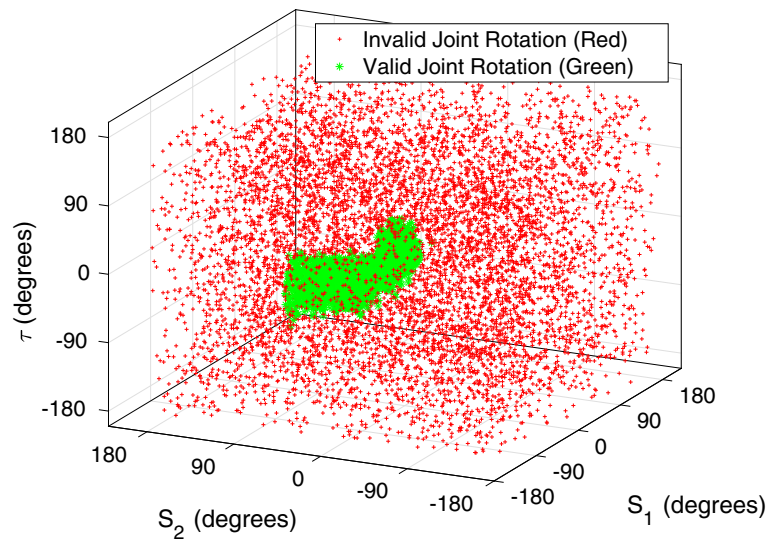
Fig. 4 Generated synthetic dataset for right hip joint with valid (green) and invalid (red) samples

number of valid/invalid instances in each batch. We train an ensemble of 20 discriminators for each joint to evaluate performance.

We also qualitatively evaluate the ability of neural network to discriminate valid/invalid joint rotations by means of an interactive validation tool. The tool also helps us to ascertain any anomalous examples found in motion capture datasets contradictory to the learned joint limits.

5   Results and discussion

The learning of a smooth implicit function by neural network represents the valid range of rotation for a specific joint, i.e, valid swing and twist component of joint rotation.

We achieve over 95% accuracy (on all joints discriminators) using only one hidden layer in a fully connected neural network (FCN) (results are shown in Table 1). The table reports the result obtained for best among 20 classifier trained for each joint. The number of optimal hidden nodes obtaining best result is also reported for each joint in Table 1. A choice of 4, 8, 16, and 32 hidden nodes is tested. The optimal number of hidden nodes is decided when performance plateau is reached. The network with eight hidden nodes is found optimal for most of the joints.

In our view, this is only feasible due to our choice of joint rotation parameterization, which is free of ambiguities and discontinuities over the joint rotation manifold. The discretization scheme proposed on swing map (s_1, s_2) for complete range of τ rotation is also more intuitive. It avoids the limitations mentioned by [15] in his earlier work on quaternion-based joint limits. The representation of joint manifold in quaternion space and Euler angle space is further highlighted in Figs. 5 and 6 respectively.

The anomalies of both representations can be noted versus our proposed representation shown in Fig. 4. The discretization of joint space in quaternion and Euler angles to label invalid region is not straightforward as evident from these figures. Therefore, no further analysis is offered in this paper.

The output of discriminator also shows that learning is smooth and effectively accounts for outliers in the data. The corresponding outliers in *synthetic* dataset which are learned as invalid by discriminator are shown in Fig. 7.

Table 1 The joint discriminator networks classification accuracy on synthetic datasets of respective joint rotations

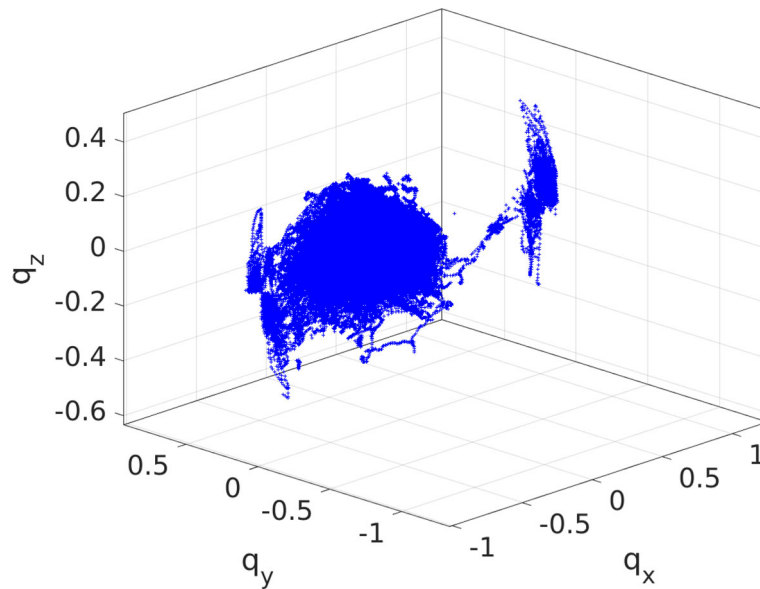| Joint name [16]	Accuracy (%)	False positive (%)	False negative (%)	Optimal hidden layer
Right up leg	98.7	0.3	1.0	8
Right foot	97.7	1	1.3	8
Left up leg	98.9	0.3	0.8	8
Left foot	98.2	0.6	1.2	8
Spine	99.4	0.1	0.5	16
Spine1	99.4	0	0.6	16
Neck	94.4	2.3	3.3	8
Head	97.7	0.9	1.4	8
Left shoulder	97.9	0.7	1.5	16
Left arm	96.8	0.5	2.7	8
Left hand	94.7	2.1	3.3	8
Right shoulder	95.1	1.5	3.3	16
Right arm	95.5	1.63	2.9	8
Right hand	95	2	3	8

Fig. 5 Quaternion (vector part) joint rotation space of right hip (right-up-leg) used by Herda et al.[15] shows duality of valid joint rotation. Anti-podal points + q and − q in the cluster represent same rotation

However, the accuracy on 'synthetic' ground truth is not an absolute figure of merit, due to voids and sparse data. Hence, the precision of joint rotation classification in a given pose is further evaluated qualitatively using an interactive tool. The output of our interactive pose validation tool is as depicted in Fig. 8. There was no improvement found in precision (vs accuracy trade-off) as well with further increase in number of neurons or hidden layers for joint discriminators. As a pilot study, we perform only limited qualitative tests to check veracity of our method. The detailed evaluation with multiple participants to validate synthetic invalid poses and evaluate discriminator performance will be undertaken in extended work.

The *intra-joint* (parent-child) dependency of range of rotation in a hierarchy of joints [9, 15] is implicitly learned
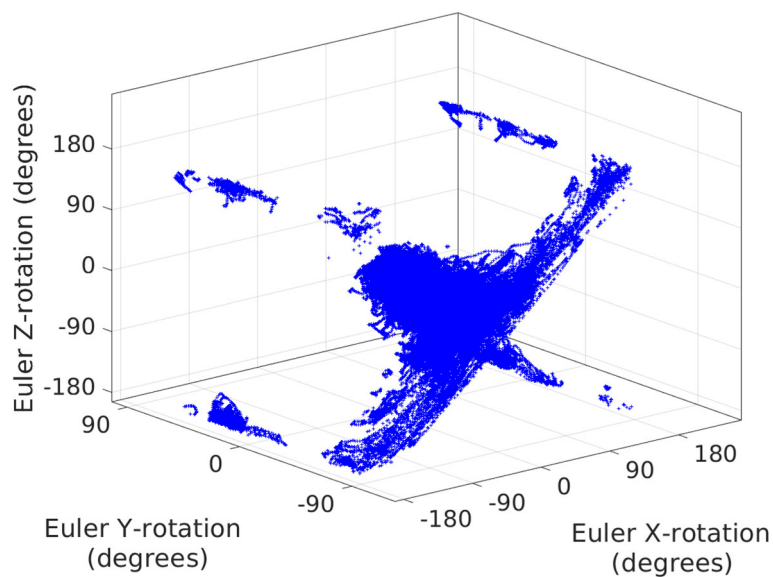
Fig. 6 The artifacts in Euler angle space of right hip (right-up-leg) shows non-contiguous clusters of valid joint rotation. The boundary jumps at 180° (for *Z* and *X*) and 90° for *Y* rotation can also be seen
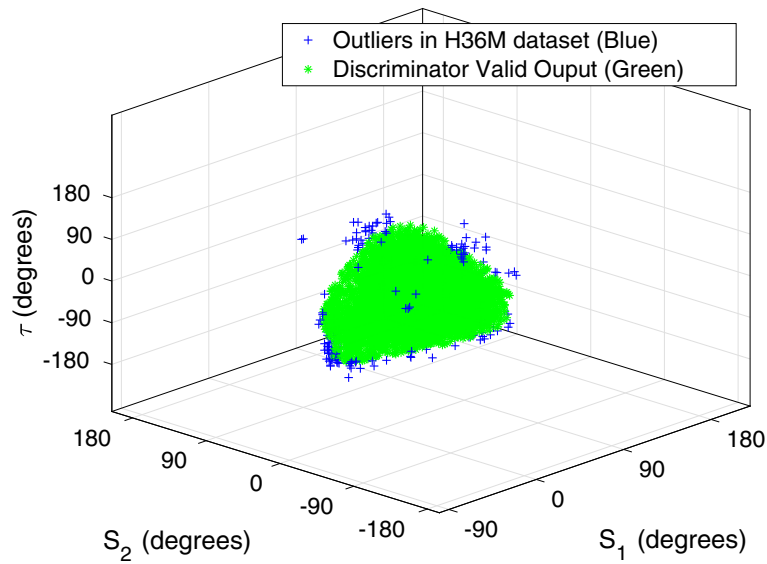
Fig. 7 Discriminator network output of valid left shoulder joint rotations in (s_1, s_2, τ) space

in our approach. Our approach emphasizes the coupling of parent-child arising from twist itself. For example, a constraint on the twist of upper limb avoids *unrealistic* pose of the lower limb by twist propagation along the kinematic chain. Thus, presented with any (s_1, s_2, τ) parameterized joint rotation, our trained neural network(s) are able to decide for the valid or invalid joints independently.

The Table 2 shows a discriminative comparison matrix on test poses of left shoulder using Akhter et al. [9] and our

trained network. All poses classified as valid by our network are also found valid by [9]. However, since [9] does not take into account twist, the joint rotations classified as invalid by our network on account of twist are still valid in [9]. Figure 9 further highlights that Akhter et al. isValid() function allows unrealistic swing of left shoulder as almost whole (90%) of swing space is classified as valid. On the contrary, valid region in swing space learned by our discriminator is more plausible. These two factors account
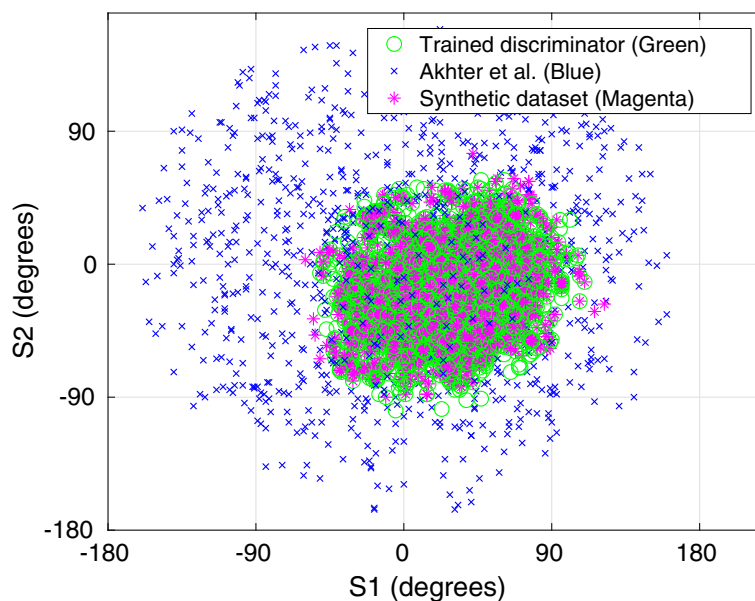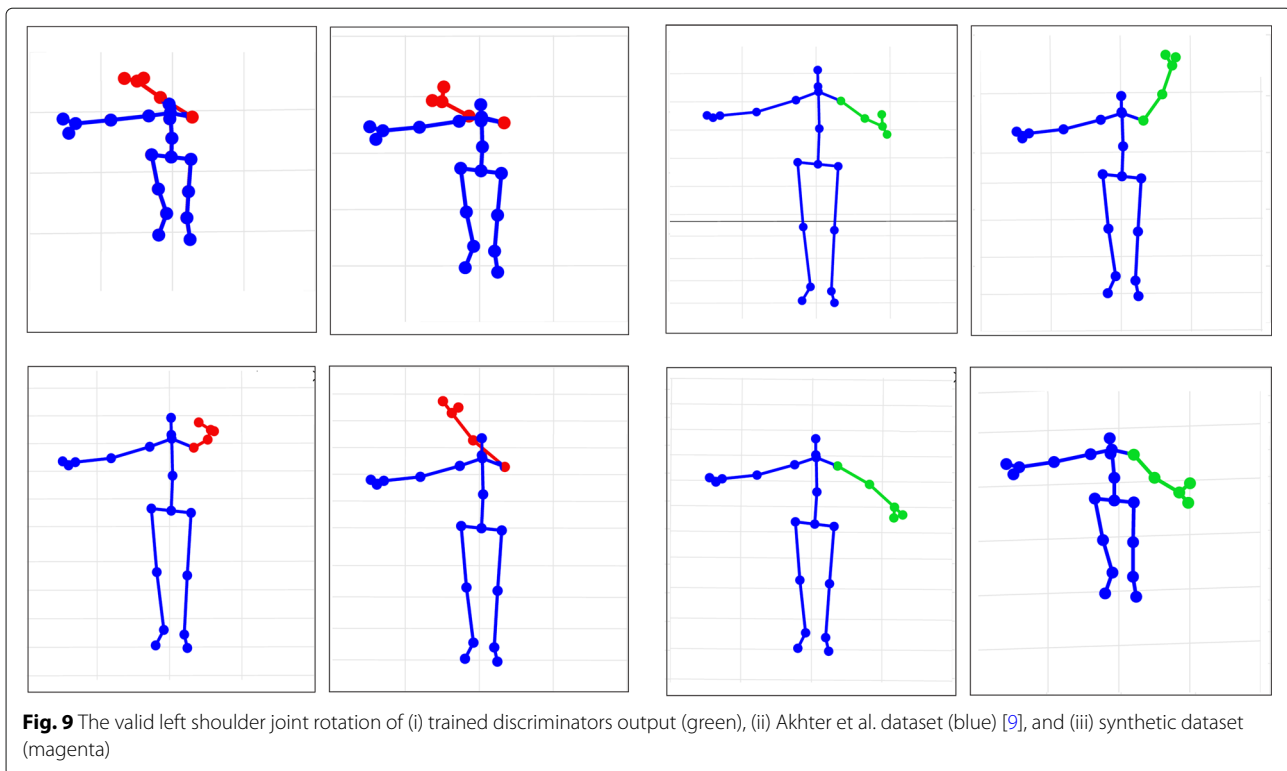

Fig. 8 Interactive validation tool output showing invalid left arm poses (in red) and valid left arm poses (in green)

Fig. 9 The valid left shoulder joint rotation of (i) trained discriminators output (green), (ii) Akhter et al. dataset (blue) [9], and (iii) synthetic dataset (magenta)

for high instances (36.65%) of left shoulder poses classified as invalid by our network while still being classified as valid by [9] in Table 2.

6 Conclusion

The articulated movement of human body is constrained by the limits on 3D joint rotations. We present a formulation of swing and twist components of joint rotations in exponential maps to learn joint limits by data-driven approach. Our representation is used to encode valid poses in an intuitive and physically meaningful way. A discretized look-up map is derived to label valid/invalid rotations in parametrized joint space. And we train a set of neural network discriminators to classify valid and invalid rotations for different joints. Our trained discriminators used as priors will lead to accurate learning of 3D human pose estimation and motion synthesis. The learned joint constraints are useful to represent normal range of human motion especially vital for social affordance applications like vision-based human-robot interaction. Data-driven approach also provides meaningful limits to monitor the rehabilitation of patient's with joint injuries. Since the learning is performed using a data-driven approach, the quality of learned constraints depends on the range of motion available in the mocap data. Therefore, we would be extending the current method by capturing more data and using other publicly available motion capture datasets. Our visualization tool can guide collection of motion data close to constraints. A more comprehensive qualitative study to validate synthetic invalid poses with the help of human evaluation would also result in better ground truth generation.

Endnote

[1] Inter-joint dependency is defined between the parent joint and the subsequent children joint angles in a kinematic chain, while intra-joint dependency exists between swing and twist of the same joint.

Table 2 Comparison of classification results between learned discriminator (ours) and Akhter et al. [9] on generated left shoulder synthetic dataset

		Akhter et.al	
		Valid	Invalid
Ours	Valid	*99.92* %	0.08%
	Invalid	*36.65* %	63.35%

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Augmented Vision, Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Trippstadter Straße 122, 67663 Kaiserslautern, Germany. [2]TU Kaiserslautern, Kaiserslautern, Germany. [3]National University of Sciences & Technology, Islamabad, Pakistan.

References

1. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Regognition (CVPR): 19-21 June 2018; Salt Lake City Utah, USA. IEEE. pp 7122–7131
2. Mehta D, Sridhar S, Sotnychenko O, Rhodin H, Shafiei M, Seidel H-P, Xu W, Casas D, Theobalt C (2017) Vnect: Real-time 3d human pose estimation with a single RGB camera. ACM Trans Graph (TOG) 36(4):44
3. Zhou X, Sun X, Zhang W, Liang S, Wei Y (2016) Deep kinematic pose regression. In: European Conference on Computer Vision (ECCV): 8-16 October 2016; Amsterdam, Netherlands. Springer, Cham. pp 186–201
4. Park S, Hwang J, Kwak N (2016) 3d human pose estimation using convolutional neural networks with 2d pose information. In: European Conference on Computer Vision (ECCV): 8-16 October 2016; Amsterdam, Netherlands. Springer, Cham. pp 156–169
5. Tekin B, Marquez Neila P, Salzmann M, Fua P (2017) Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV): 22-29 October 2017; Venice, Italy. IEEE. pp 3941–3950
6. Varol G, Romero J, Martin X, Mahmood N, Black MJ, Laptev I, Schmid C (2017) Learning from synthetic humans. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 21-26 July 2017; Hawaii, United States:109–117. IEEE
7. Pavllo D, Grangier D, Auli M (2018) QuaterNet: A Quaternion-based Recurrent Model for Human Motion. In: British Machine Vision Conference (BMVC): 3-6 September 2018; Newcastle, UK. BMVA. p 299
8. Elhayek A, Kovalenko O, Murthy P, Malik J, Stricker D (2018) Fully automatic multi-person human motion capture for vr applications. In: EuroVR International Conference on Virtual Reality and Augmented Reality: 22–23 October 2018; London, UK. Springer, Cham. pp 28–47
9. Akhter I, Black MJ (2015) Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Regognition (CVPR): 8-10 June 2015; Boston Massachusetts, USA. IEEE. pp 1446–1455
10. Engell-Nørregård M, Erleben K (2009) Estimation of joint types and joint limits from motion capture data. In: International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG): 2-5 February 2009; Plzen, Czech Republic. Eurographics. pp 9–16
11. Baerlocher P, Boulic R (2001) Parametrization and range of motion of the ball-and-socket joint. In: Thalmann NM, Thalmann D (eds). Deformable Avatars. Springer. pp 180–190
12. Shao W, Ng-Thow-Hing V (2003) A general joint component framework for realistic articulation in human characters. In: Proceedings of the 2003 symposium on Interactive 3D graphics (I3D): 27 - 30 April 2003; Monterey California, USA. ACM. pp 11–18
13. Dobrowolski P (2015) Swing-twist decomposition in clifford algebra. arXiv: Computing Research Repository (CoRR) 1506.05481. https://arxiv.org/abs/1506.05481v1
14. Grassia FS (1998) Practical parameterization of rotations using the exponential map. J Graph Tools 3(3):29–48
15. Herda L, Urtasun R, Fua P, Hanson A (2003) Automatic determination of shoulder joint limits using quaternion field boundaries. Int J Robot Res 22(6):419–436
16. Ionescu C, Papava D, Olaru V, Sminchisescu C (2014) Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell 36(7):1325–1339