SN Computer Science (2023) 4:760
https://doi.org/10.1007/s42979-023-02195-0

ORIGINAL RESEARCH

Neural Network-Based Human Motion Predictor and Smoother

Stella Graßhof[1] · Mathias Bastholm[2] · Sami S. Brandt[1]

Received: 20 May 2022 / Accepted: 28 July 2023

Abstract
Though continuous advances in the field of human pose estimation, it remains a challenge to retrieve high-quality recordings from real-life human motion using commodity hardware. Therefore, this work focuses on predicting and improving estimates for human motion with the aim of achieving production quality for skinned mesh animations by off-the-shelf webcams. We take advantage of recent findings in the field by employing a recurrent neural network architecture to (1) predict and (2) denoise human motion, with the intention of bridging the gap between cheap recording methods and high-quality recording. First, we propose an LSTM to predict short-term human motion, which achieves competitive results to state-of-the-art methods. Then, we adapt this model architecture and train it to clean up noisy human motion from two 3D low-quality input sources, and hence mimic a real-world scenario of recording human motion which yields noisy estimates. Experiments on simulated data show that the model is capable of significantly reducing noise, and it opens the way for future work to test the model on annotated data.

Keywords Recurrent neural networks · Reconstruction · Computer vision · Animation denoising

Introduction

Human motion and psychology are interconnected, as movements reflect and express emotions, contribute to cognitive development, serve as nonverbal communication, and are used in therapeutic applications, highlighting the close relationship between the mind and the body.

Human motion can be represented as a sequence of 3D joints connected via lines representing segments, see Fig. 1, or as a sequence of angles between the segments which we describe in more detail in Section "Human Motion Parameterisation.

The creation of a realistic, human motion animation as a skinned mesh animation is difficult with the production quality. The skinned multi-person linear (SMPL) models [2–5] express the pose and shape of human bodies in a sparse manner. This is accomplished by representing the human as a skinned mesh, with blend shapes representing the shape of the human, and the underlying skeleton of the skinned mesh representing the pose. Having chosen one representation of human motion, new samples can be generated using neural networks based on different kinds of input [6–11].

3D meshes of the human body are usually built around a skeleton for the purpose of animating the human motion. These skeletons are then either animated by hand or by using motion capture (MoCap) to capture real-life human motion as digital animations. Animating skeletons by hand is a time-consuming process and requires a skilled animator. Likewise, MoCap requires specialised equipment and often also requires an animator to clean up the recorded data. Animating humanoids thus consumes a lot of time and money for content creators.

Recently, several solutions allowing for MoCap from a single video camera have been published [12–15]. These are

This article is part of the topical collection "Advances on Pattern Recognition Applications and Methods 2022" guest edited by Ana Fred, Maria De Marsico and Gabriella Sanniti di Baja.

The conference version of this work is [1].

✉ Stella Graßhof
stgr@itu.dk

Mathias Bastholm
mathias@marionettexr.com

Sami S. Brandt
sambr@itu.dk

[1] Computer Science Department, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen, Denmark

[2] Marionette ApS, Copenhagen, Denmark

Published online: 29 September 2023

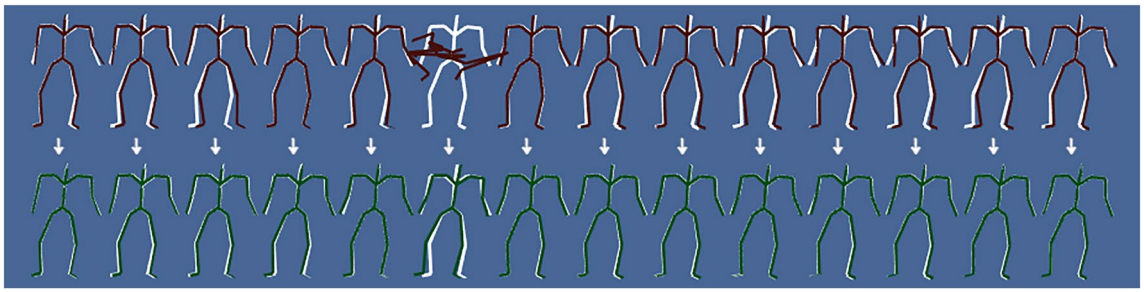

Fig. 1 In the top row, the input data is shown in red, the bottom row shows the output of our model in green, and the ground truth is in white

not widely used, which is likely because the quality is much lower than that of MoCap and handcrafted animations. They would require a significant cleanup pass by an animator to be of use, even for projects with relatively low animation quality requirements.

In this paper, we propose using recent advances in the prediction of human motion through neural networks to improve the quality of human motion, to bridge the gap between cheap recording methods and high-quality recording, see Fig. 1. The main novelty of this paper, however, is the proposition to use two inexpensive, low-quality sources for human pose in 3D and the smoothing of them to achieve high-quality human motion sequences. In other words, the model is trained to clean up 3D human motion from two low-quality input recordings.

To conclude, the contributions of this work are the following.

- We modified the short-term version of QuaterNet [16] by
  - using a long short-term memory (LSTM) network instead of a gated recurrent unit (GRU) network, and,
  - redefining the loss function as the L1 distance between the predicted and ground truth quaternions.
- We are the first to propose a model which receives two noisy 3D human motion sequences to perform a de-noising, which promotes a cost-efficient imaging solution, e.g. using webcams.
- We show that the same architecture can be used for two separate tasks: (1) prediction and (2) smoothing of human motion.

This paper is organised as follows. First, the related work is described in Section "Related Work". Then prior work on predicting human motion is replicated and extended to the task of motion smoothing in Section "Methods". The results of the proposed models on both prediction and smoothing of human motion can then be seen in Section "Experiments", followed by a section dedicated to limitations in "Limitations". Finally, we conclude this paper with a thorough discussion and mention the proposed future work in Section "Conclusion".

Related Work

Human Motion Prediction

Forecasting human motion is an important problem in computer vision. It is a central problem, in addition to creating digital animations of human motion, in applications like human robot interaction, autonomous driving, and human tracking. The problem is challenging due to the high variability and the complex nature of the human motion. Traditional, state–space methods, such as hidden Markov models [17] and Gaussian processes [18], have been shown to be suitable for the prediction of simple human motion; see Table 1 for an overview of the methods used for human motion prediction.

In the literature, see e.g. [16, 19], the prediction of sequences of 3D joint positions is commonly divided into short- and long-term predictions. Specifically, *short-term* refers to predictions limited to under 500 ms, while *long-term* tasks focus on motions which lay more than 0.5 s in the future, hence often referred to as *generation*. Recently deep neural networks, especially recurrent neural networks (RNNs), have made larger advances in the prediction of human motion for a longer range [16, 20–23].

QuaterNet [16], as proposed by Pavllo et al., consists of a two-layer RNN predicting future human motion from past motion, using a forwards kinematics (FK) loss. In that work, the rotations are represented by quaternions, opposed to previous works, where Euler angles or exponential maps are frequently employed. The choice was motivated by the fact that Euler angles and axis-angle representations come with several problems: non-uniqueness, discontinuity in the representation space, and singularities,

Table 1 Overview of related literature for human motion prediction

References	Architecture	Notable merits/limitations
Wang et al. [18]	Gaussian Processes	Performance limitations due to time steps being correlated
QuaterNet [16]	RNN, two-layer GRU	Locomotion only
Wolter & Yao [23]	Complex gated RNN	Prediction error accumulation
Li et al. [24]	Multi-scale graph conv.	Efficient formulation
Dang et al. [25]	Multi-scale graph conv.	Additional residual connections
Mao et al. [27]	Graph conv.	Trajectory space formulation
Mao et al. [26]	Graph conv.	Attention
Liang et al. [32]	LSTM, NN	Simultaneous prediction of trajectory and future activity
Yuan and Kitani [42]	VAE	Diversity-promoting prior
Cao et al. [28]	VAE, GAN	Uses scene geometry constraints
Ruiz et al. [30]	GAN	Presented metrics not comparable
Amirian et al. [31]	LSTM, InfoGAN	Limited to one motion trajectory per person
Martínez et al. [43]	Transformer	A single pose vector as query

which can be avoided by quaternions. QuaterNet also introduces a normalisation loss, as normalised quaternions are required to represent valid rotations. The FK loss is calculated by performing FK and then taking the positional loss of the joints. FK is when the joint positions are calculated from the joint rotations using the pre-defined skeleton. FK loss helps against the positional error introduced on the outer limbs by rotational error on the inner limbs, as the positional error of the outer limbs is affected by the rotational error of all parent limbs in the kinematic chain.

Another branch of human motion prediction networks are based on graph representation of the human body and related graph computations. Li et al. [24] proposed a multi-scale graph representation of the human body and encoder–decoder framework for motion prediction. An alternative, end-to-end multi-scale residual graph convolution network was proposed in [25]. Mao et al. [26] proposed motion attention to a graph convolutional network to capture the similarity between the current motion context and the historical motion sub-sequences. To encode temporal information, trajectory space was applied instead of the traditionally used pose space in [27].

Deep generative models, variational autoencoders (VAEs) and generative adversarial networks (GANs) have also been used for human motion prediction with the special aim of facilitating human motion prediction for the long horizon [28]. As an example, the conditional variational autoencoder (CVAE) was used to generate a diverse set of samples of human postures from a pretrained deep generative model in [29]. Spatio-temporal motion inpainting was proposed by a GAN prediction model in [30] and pedestrian trajectories were learnt with GANs in [31]. One more example of long-term human activity and location prediction was proposed in [32].

Human Motion Inpainting

Harvey et al. [33] showed that state-of-the-art motion prediction models cannot be easily converted into a robust transition generator, and proposed a model for human motion inpainting, i.e. a method that can fill in gaps of missing motion in a given motion sequence. It takes past motion and a target frame as input and then generates the frames in between using an RNN. To help the model maintain temporal coherency, a time-to-arrival embedding was added to the input frames.

To create realistic looking and temporally coherent motion, an adversarial loss based on least squares generative adversarial network [34] (LSGAN) was introduced.

Additionally, [33] uses a foot contact loss indicating whether a foot is touching the ground, thereby stabilising the feet as a post-processing step, which helps to combat a phenomenon commonly known as *foot sliding*. Foot loss can also be found in another recent work involving human motion, such as MotioNet [14].

Human Motion Smoothing

In the previously described work, the application focused on prediction or inpainting, which mostly relies on reliable estimates of human motion data as a starting point. However, raw motion data is often corrupted, i.e. the markers attached to the joints may be occluded, or lack precision, and hence yield noisy and jittery estimates or even miss data entirely.

Table 2 Overview of related literature for human motion smoothing and de-noising

Reference	Architecture
Lou et al. [35]	Sequence filtering
Piltaver et al. [36]	Kalman filter
Brand and Hertzmann [17]	HMM
Dagioglou et al. [37]	Bezier curves
Memar et al. [38]	B-spline-based least squares
Kim et al. [39]	Bidirectional recurrent neural network (BRNN)
Cui et al. [40]	Deep bidirectional attention network (BAN)

To overcome these issues, research has been conducted to smooth and denoise human motion data; see Table 2.

As to human motion smoothing methods, in [35] used traditional filtering methods and [36] proposed Kalman filtering. While these are older works, we found that traditional methods are still used these days, e.g. in [37] Bezier curves are used and [38] achieves a significant noise reduction by a B-spline-based least squares approach on data from a vicon motion capture system. Different network-based approaches have been employed to tackle the problem of corrupted data. In [39], an attention-based bidirectional recurrent neural network [39] was proposed to denoise hand motion data. Similarly, in [40], an attention mechanism was embedded in the bidirectional LSTM (BLSTM) yielding a deep bidirectional attention network (BAN). However, the current approaches for smoothing do not consider multiple input sources which is a shortcoming we address in this paper. We propose using two corrupted input sequences of 3D human motion to retrieve a smoothed version of the recorded motion. Our motivation for this approach is that currently several solutions to retrieve estimates for 3D human motion sequences are available, e.g. in [41], which come with some errors. We aim to take advantage of several of those corrupted 3D estimates to retrieve one high-quality sequence of human motion.

Methods

In this section, we first introduce two models which are designed to perform two separate tasks. First, we describe a prediction model, which receives past frames of human motion to predict the next frame one step in the future.

Second, we adapt the prediction model such that it will be tailored to the task of denoising human motion data aka human motion smoothing. Both models are builton an RNN architecture.

Human Motion Parameterisation

The human skeleton applied in this work is parameterised as follows. The joint locations are represented by 3D joint positions where joints are connected to other joints by line segments. If the joint positions are estimated for each frame separately, the length of the segments may change between frames, which is a common problem [44]. However, assuming constant lengths of the line segments, the configuration of the human skeleton can be fully defined by the relative orientations of the line segments, where each orientation is described by the respective quaternion, i.e. 3D angle.

The use of quaternions avoids the gimbal lock problem present with Euler angles.

Prediction Model

The proposed prediction model is designed to use past frames of a human motion sequence to predict the deterministic motion in the next frame. The proposed model is based on the short-erm version of QuaterNet [16], which we modified in two ways: frst, motivated by results from Harvey et al. [33], we use a long short-term memory (LSTM) network instead of a gated recurrent unit (GRU) network. Secondly, we adapted the rotational loss. Instead of using the distance between Euler angles calculated from quaternions,we redefine it as the L1 distance between the icted and ground truth quaternions as

$$L_{\text{prediction}} = \frac{1}{T}\sum_{t=0}^{T}\sum_{j=0}^{J}\left\|\hat{\mathbf{q}}_{j,t} - \mathbf{q}_{j,t}\right\|_1, \tag{1}$$

where T is the sequence length, J is the number of considered joint rotations of the skeleton, $\hat{\mathbf{q}}_{j,t}$ is the joint rotation j of the predicted sequence at time step t, and $\mathbf{q}_{j,t}$ the corresponding ground truth, both represented as quaternions. Hence, we combine the rotational error and the quaternion normalisation error by dropping the explicit term for normalisation used in [16].

The prediction model is designed as an encoder–decoder LSTM, with a two-ayer LSTM encoder with a hidden state

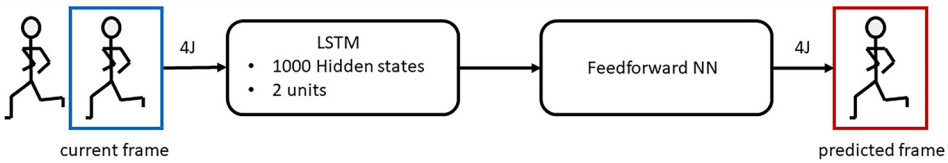

Fig. 2 The architecture of the prediction model

of size 1000, followed by a feedforward neural network as the decoder. The decoder converts the hidden state to the target output, size of $4J$, see Fig. 2. The model receives the past 50 frames to predict the next frame. To train the model, this process is repeated ten times to generate the next ten frames from the previous model outputs.

Smoothing Model

The process of capturing 3D human motion is not exact and yields noisy estimates. To overcome this issue, we aim to provide one noise-free estimate from two noisy recordings of the same motion. For this purpose, to remove the stochastic part of the motion, a *smoothing model* is designed based on the previously introduced prediction model, i.e. an encoder–decoder LSTM, where the encoder part is defined as a two-layer LSTM with hidden state size of 1000, and the decoder as a feedforward neural network. This model uses the same loss function as the prediction model, as defined in Eq. 1.

In contrast to the prediction model, the smoothing model receives two concatenated frames per time step as input, thereby reflecting the real-world setting that two noisy input streams are provided, e.g. from different camera angles.

From each pair of noisy frames, the network estimates one corresponding noise-free frame. See Section "Data Generation for the Smoothing Model" for details on how the training data were generated for this model.

Experiments

Datasets

In this work, we use two 3D human motion capture datasets: the CMU [45, 46] and the Human3.6M dataset [47, 48].

The CMU MoCap Dataset [45] consists of 2605 human motions of 106 subjects recorded in 3D, totaling 552 minutes of motions at varying frame rates and $3.5M$ frames. We use the skeleton model which is fully parameterised by 22 body segment orientations.

The CMU Mocap Dataset is one of the 15 optical marker-based MoCap datasets which have been represented in a common framework and parameterisation in the Archive of Motion Capture as Surface Shapes (AMASS) database [46]. Specifically, an SMPL-H [3] variant of the *Skinned Multi-person Linear Model* (SMPL) [2] was used.

The Human3.6M dataset [47, 48] contains over 3.6 million different human poses, recorded in 2D and 3D from seven subjects performing 210 different motions in 15 subcategories. This yields a total of 176 min recorded at 50 frames per second and 0.5M frames. While the database also contains high-resolution 3D meshes, we focus on the sparse skeleton which consists of a total of 32 joints, and for which 3D joint positions and joint angles are provided.

Implementation Details

We trained the models as follows. We used Adam as optimiser with a learning rate of 0.001, and the gradient norms are clipped to 0.1. Training data is batched, with a batch size of 64. The batches are drawn from the dataset by taking all possible combinations of 60 consecutive frames for each motion and shuffled for each epoch. Additionally, we used teacher forcing to improve the prediction and decrease the training time. We trained the models until there was very little or no improvement, which took several days, close to a week.

Prediction Model

The prediction model proposed in Section "Prediction Model" was trained on the Human3.6M dataset; see Section "Datasets". We split the dataset as in [16, 20, 33] by using all the motions from subject 5 as test data and the rest as training data. Additionally, since the frame rate differs between the motions, we resampled them to 25 fps by either discarding frames or interpolating new frames, i.e. by down- or upsampling the sequence, respectively. The prediction is performed by taking 60 frames from a motion and then splitting it into 50 past and 10 future frames.

The results were evaluated using the mean absolute error between the Euler angles as

Table 3 Comparison of the mean absolute error between Euler angles, as defined in Eq. 2, between our proposed prediction model and other state-of-the-art methods on the Human3.6M dataset

Milliseconds	Walking				Eating				Smoking				Discussion			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
Zero-velocity [20]	0.39	0.68	0.99	1.15	0.27	0.48	0.73	0.86	0.26	0.48	0.97	0.95	0.31	0.67	0.94	1.04
QuaterNet [16]	0.21	**0.34**	**0.56**	**0.62**	0.20	0.35	0.58	0.70	0.25	0.47	0.93	0.90	0.26	0.60	**0.85**	**0.93**
TP-RNN [21]	0.25	0.41	0.58	0.65	0.20	**0.33**	**0.53**	0.67	0.26	0.47	0.88	0.90	0.30	0.66	0.96	1.04
ERD-QV [33]	**0.20**	**0.34**	0.56	0.64	**0.18**	**0.33**	**0.53**	**0.63**	0.23	0.47	0.96	0.99	**0.23**	**0.59**	0.86	**0.93**
VGRU-rl [22]	0.34	0.47	0.64	0.72	0.27	0.40	0.64	0.79	0.36	0.61	**0.85**	0.92	0.46	0.82	0.95	1.21
Our model	0.24	0.40	0.61	0.68	0.21	0.37	0.57	0.69	**0.23**	**0.44**	0.89	**0.88**	0.26	0.64	0.93	1.00

Bold values indicate the best result per column

$$L_{\text{mae}} = \frac{1}{T}\sum_{t,j}\left\|\left(\Phi\left(\hat{\mathbf{q}}_{t,j}\right) - \Phi\left(\mathbf{q}_{t,j}\right) + \pi\right) \bmod 2\pi - \pi\right\|_1, \quad (2)$$

where T is the sequence length, J is the number of joint rotations, $\hat{\mathbf{q}}_{t,j}$ represents the predicted quaternion of the joint j in frame t, with corresponding ground truth $\mathbf{q}_{t,j}$, and Φ is a function converting quaternions into Euler angles. In Table 3, we compare our results with others, which show that our proposed prediction model is on par with state-of-the-art methods.

Data Generation for the Smoothing Model

To our knowledge, there is no dataset which offers noisy human motion capture data, along with a non-noisy ground truth; therefore, we created our own data based on the CMU dataset Section "Datasets" to train the smoothing model, described in Section "Smoothing Model". We employ the provided data as ground truth frames and create noisy input frames from them to train and evaluate our model. Therefore, 60 frames from one motion are selected and then split into 50 past and 10 future frames. Thereafter, noise is added to all frames, i.e. input features, as

$$\tilde{\mathbf{q}}_{m,t,j,a} = \mathbf{q}_{m,t,j,a} + N, \quad (3)$$

where m is the human motion, t is the frame, j is the joint, a is the axis, $\mathbf{q}_{m,t,j,a}$ is the ground truth quaternion, $\tilde{\mathbf{q}}$ is the noisy quaternion, and N represents the noise.

When modelling the noise N, we take into account three different kinds of noise: systematic bias, imprecision, and lost tracking. The noise N is defined as a composition of those as

$$N = B + I + L, \quad (4)$$

where B is the bias, I is imprecision noise, and L represents the noise from lost tracks.

It has been observed that joint rotations captured through webcam pose detection systems often have a constant bias, depending on the subject captured, which is represented as

$$B \sim \mathcal{N}\left(0, \theta_B^2\right), \quad (5)$$

$$\theta_B \sim \mathcal{N}\left(\mu_B, \sigma_B^2\right), \quad (6)$$

where $\mathcal{N}$ denotes the normal distribution. The imprecision noise represents small differences from the ground truth that occur in the joint rotations captured through webcam pose detection models as

$$I \sim \mathcal{N}\left(0, \theta_I^2\right), \quad (7)$$

$$\theta_I \sim \mathcal{N}\left(\mu_I, \sigma_I^2\right). \quad (8)$$

The lost tracking noise L represents that sometimes a joint is not recognised, giving completely arbitrary values for that joint rotation, defined as

$$L = L_1 L_2, \quad (9)$$

where

$$L_1 \sim \mathcal{B}\left(p_L\right) \quad (10)$$

models the probability that the model lost track of one frame, where $\mathcal{B}$ denotes the Bernoulli distribution, and

$$L_2 \sim \mathcal{N}(0, \theta_L^2), \quad (11)$$

$$\theta_L \sim \mathcal{N}(\mu_L, \sigma_L^2) \quad (12)$$

models the amount of noise which is applied if the frame suffers from lost tracking. To conclude, to define the noise, we use a total of seven parameters.

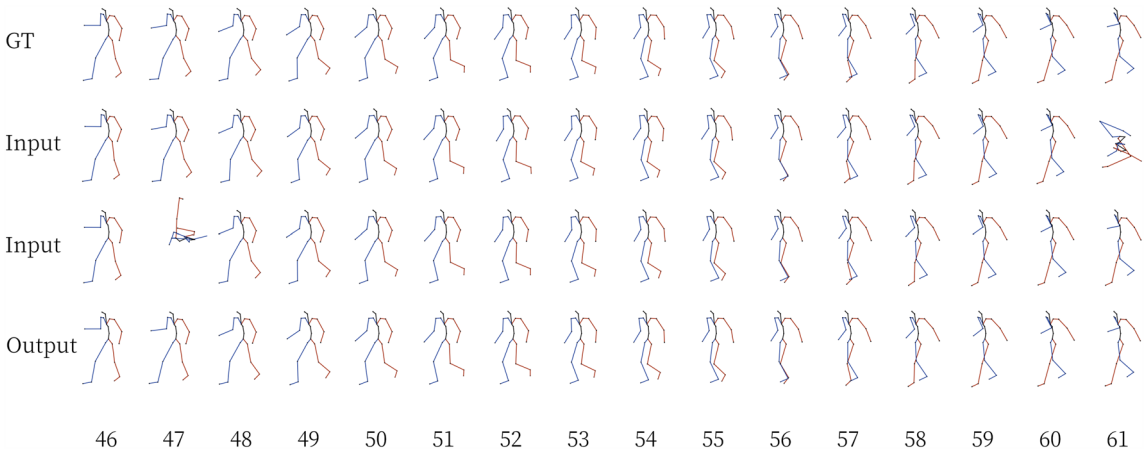

Fig. 3 Visualisation of 15 selected frames of subject 6, trial 5 from the CMU dataset. The first row shows the ground truth poses, while the second and third row show the noisy frames generated from the GT, which are concatenated and fed to the model. The fourth row shows the resulting output from the smoothing model. Please especially note the results of the frames 47 and 61, which demonstrate that the model is robust against lost frames. A video of the results is available at https://i.imgur.com/gS3Pin8.mp4

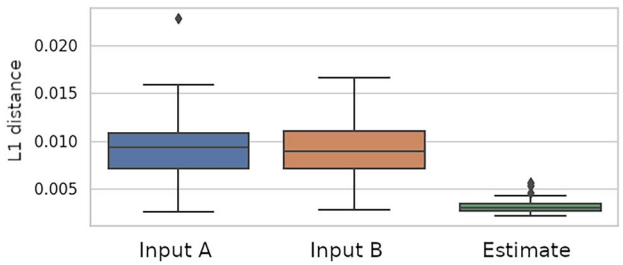

Fig. 4 Comparison of the L1 distance between quaternions of the ground truth data (GT) and noisy input A (shown in blue), the GT and the noisy input B (shown in red), and the GT and our estimates (shown in green). Inputs A and B refer to the two concatenated views, which the smoothing model receives as input

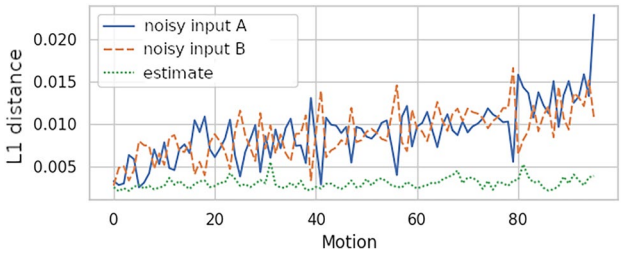

Fig. 5 Illustration of the L1 distance between estimated quaternions to the ground truth (GT) values per motion of the test data generated from the CMU dataset. For convenience, the motion categories are ordered along the *x*-axis such that the average distance of the two noisy input sequences A and B to the GT increases

Smoothing Model

The smoothing model, described in Section "Smoothing Model", was trained and evaluated on the CMU dataset, see Section "Datasets", with added noise according to Section "Data Generation for the Smoothing Model. To load and manipulate the motions from the CMU dataset, we use Fairmotion [49]. Since one training step requires at least 60 frames, motions with less than 60 frames have been discarded. The data is split, such that 90% of the motions are used for training, 5% for validation, and 5% for testing. For training of the smoothing model, the size of each input source is limited for batching purposes, but all frames are used for evaluation.

The training data were generated according to Section 4.4, with $\mu_B = \mu_I = 0.005$, $\sigma_B = \sigma_I = 0.002$, $\mu_L = 1$, $\sigma_L = 0.01$, and $p = 0.01$. To ensure that each motion has a unique distribution of noise, each parameter θ is only sampled once per motion, thereby ensuring that the model learns the general composite noise instead of one specific distribution. The same procedure and noise function are used to generate the test data.

Figure 3 shows the GT, the generated noisy input and the estimates from our model, and demonstrates that the model is able to recover information loss from lost frames. For a quantitative evaluation, we computed the L1 distance between quaternions of the ground truth and the noisy inputs, and the ground truth and the estimates, accordingly. The results are shown in Fig. 4 and illustrated per motion in Fig. 5. In both figures, it can be seen that the distances are lower for the smoothing model. To summarise, we found that our proposed smoothing model is able to reduce the noise to a large extent in 3D human motion sequences, thereby confirming that an LSTM-based model is suitable to for this task.

Limitations

During our experiments, we found that training did not converge, which we overcame by hand-tuning the training parameters and trying out different activation functions. While the smoothing model successfully yields smoothed, i.e. denoised, 3D human motion sequences, we found that if we provide one non-corrupted sequence, while the second input sequence is only noise, the outcome will be a jittery human motion sequence.

Conclusion

In this work, we have proposed a novel approach to estimate the human motion by merging and enhancing data from two low-quality sources. As a building block, our work also proposed an LSTM-based prediction model for human motion which was demonstrated to be competitive with previous approaches in the field. The key advantage of our approach lies in its ability to enable low-cost imaging of human motion without the need for expensive hardware traditionally associated with motion capture.

To the best of our knowledge, no suitable dataset currently exists for cleaning, i.e. smoothing of skinned human motion that would be suitable for pose detection from webcam videos.

While training the smoothing network, the lack of a dedicated dataset is not problematic, since simulated training data can be used. However, evaluating the network presents challenges due to the susceptibility of neural networks to shortcut learning [50], where the network may learn unintended shortcuts instead of the desired generalised solution. For instance, a neural network trained to classify objects might incorrectly take the background into account, leading to mislabelling.

One potential approach to mitigate shortcut learning involves evaluating the network using data from a separate dataset that was not used for training purposes. Our evaluation data is related to the training data in two ways. Firstly, the evaluation data stems from the same dataset, making it i.i.d. with respect to the motions it contains. Secondly, the noise used to generate the input motions in the evaluation is not the actual noise encountered in a webcam-based pose estimation pipeline, but rather the same noise estimation used during network training. Thus, the validation data represents the best possible effort considering the limited availability of data for this specific task. However, in the event that datasets of skinned human motion smoothing become accessible, it would be desirable to re-evaluate the model on these out-of-distribution datasets.

Consequently, the lack of annotated data for evaluation implies that the performance of the model on real-world data is uncertain. Overcoming this limitation and implementing various potential improvements is an interesting topic for future work.

Funding Open access funding provided by IT University of Copenhagen.

Data availability In this work we used the 2 datasets which were described in section "Datasets". The CMU MoCap Dataset [45] is available at http://mocap.cs.cmu.edu/, and can be copied, modified, or redistributed without permission. The second dataset which we used is the Human3.6M dataset [47]. Access to the data is subject to approval and a data sharing agreement, and will only be granted for academic use. To get acces to the data please visit http://vision.imar.ro/human3.6m/.

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Bastholm M, Graßhof S, Brandt SS. Neural network-based human motion smoother. In: Proceedings of the 11th International Conference on Pattern Recognition Applications and Methods (ICPRAM), 2022; pp. 24–30. INSTICC. https://doi.org/10.5220/0010790500003122.
2. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: a skinned multi-person linear model. In: ACM Trans. Graphics (Proc. SIGGRAPH Asia) 2015;34(6):248–124816.
3. Romero J, Tzionas D, Black MJ. Embodied hands: modeling and capturing hands and bodies together. In: ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) 2017;36(6).
4. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AAA, Tzionas D, Black MJ. Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings IEEE Conf. on computer vision and pattern recognition (CVPR), 2019; pp. 10975–10985.
5. Osman AAA, Bolkart T, Black MJ. STAR: A sparse trained articulated human body regressor. In: European Conference on Computer Vision (ECCV), 2020; pp. 598–613. https://star.is.tue.mpg.de. Accessed 28 July 2023.
6. Holden D, Komura T, Saito J. Phase-functioned neural networks for character control. ACM Trans Graph. 2017. https://doi.org/10.1145/3072959.3073663.

7. Zhang H, Starke S, Komura T, Saito J. Mode-adaptive neural networks for quadruped motion control. ACM Trans Graph. 2018. https://doi.org/10.1145/3197517.3201366.
8. Starke S, Zhang H, Komura T, Saito J. Neural state machine for character-scene interactions. ACM Trans Graph. 2019. https://doi.org/10.1145/3355089.3356505.
9. Starke S, Zhao Y, Komura T, Zaman K. Local motion phases for learning multi-contact character movements. ACM Trans Graph. 2020. https://doi.org/10.1145/3386569.3392450.
10. Ling HY, Zinno F, Cheng G, Van De Panne M. Character controllers using motion vaes. ACM Trans Graph. 2020. https://doi.org/10.1145/3386569.3392422.
11. Holden D, Kanoun O, Perepichka M, Popa T. Learned motion matching. ACM Trans Graph. 2020. https://doi.org/10.1145/3386569.3392440.
12. Rong Y, Shiratori T, Joo H. Frankmocap: a monocular 3d whole-body pose estimation system via regression and integration. In: IEEE International Conference on Computer Vision Workshops; 2021.
13. Joo H, Neverova N, Vedaldi A. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In: IEEE International Conference on 3D Vision (3DV), 2021; pp. 42–52.
14. Shi M, Aberman K, Aristidou A, Komura T, Lischinski D, Cohen-Or D, Chen B. Motionet: 3d human motion reconstruction from monocular video with skeleton consistency. ACM Trans Graph. 2020. https://doi.org/10.1145/3407659.
15. Pavllo D, Feichtenhofer C, Grangier D, Auli M. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR), 2019; pp. 7753–7762.
16. Pavllo D, Feichtenhofer C, Auli M, Grangier D. Modeling human motion with quaternion-based neural networks. Int J Comput Vis. 2019;128(4):855–72. https://doi.org/10.1007/s11263-019-01245-6.
17. Brand M, Hertzmann A. Style machines. In: Proceedings of the 27th Annual Conference on computer graphics and interactive techniques—SIGGRAPH '00, 2000; pp. 183–192. https://doi.org/10.1145/344779.344865.
18. Wang JM, Fleet DJ, Hertzmann A. Gaussian process dynamical models for human motion. IEEE Trans Pattern Anal Mach Intell. 2008;30(2):283–98. https://doi.org/10.1109/TPAMI.2007.1167.
19. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human36.M: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell. 2014;36(7):1325–39. https://doi.org/10.1109/TPAMI.2013.248. (**Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence**).
20. Martinez J, Black MJ, Romero J. On human motion prediction using recurrent neural networks. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR), 2017; pp. 4674–4683. https://doi.org/10.1109/CVPR.2017.497.
21. Chiu H-K, Adeli E, Wang B, Huang D-A, Niebles JC. Action-agnostic human pose forecasting. In: 2019 IEEE Winter Conference on applications of computer vision (WACV), 2019; pp. 1423–1432. https://doi.org/10.1109/WACV.2019.00156.
22. Gopalakrishnan A, Mali A, Kifer D, Giles L, Ororbia AG. A neural temporal model for human motion prediction. In: 2019 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), 2019; pp. 12108–12117. https://doi.org/10.1109/CVPR.2019.01239. ISSN: 2575-7075
23. Wolter M, Yao A. Complex gated recurrent neural networks. In: Proc. 32nd Conference on neural information processing systems (NeurIPS 2018), Montréal, Canada 2018.
24. Li M, Chen S, Zhao Y, Zhang Y, Wang Y, Tian Q. Dynamic multi-scale graph neural networks for 3D skeleton based human motion prediction. In: 2020 IEEE/CVF Conference on computer vision and pattern recognition (CVPR), pp. 211–220. IEEE, Seattle, WA, USA 2020. https://doi.org/10.1109/CVPR42600.2020.00029.
25. Dang L, Nie Y, Long C, Zhang Q, Li G. MSR-GCN: multi-scale residual graph convolution networks for human motion prediction. In: 2021 IEEE/CVF International Conference on computer vision (ICCV), pp. 11447–11456. IEEE, Montreal, QC, Canada. 2021. https://doi.org/10.1109/ICCV48922.2021.01127.
26. Mao W, Liu M, Salzmann M. history repeats itself: human motion prediction via motion attention. In: Vedaldi A, Bischof H, Brox T, Frahm J-M. editors. Proc. European Conference on Computer Vision—ECCV 2020, vol. 12359, pp. 474–489. Springer, Cham 2020. https://doi.org/10.1007/978-3-030-58568-6_28. Series Title: Lecture Notes in Computer Science.
27. Mao W, Liu M, Salzmann M, Li H. Learning trajectory dependencies for human motion prediction. In: 2019 IEEE/CVF International Conference on computer vision (ICCV), pp. 9488–9496. IEEE, Seoul, Korea (South) 2019. https://doi.org/10.1109/ICCV.2019.00958.
28. Cao Z, Gao H, Mangalam K, Cai Q-Z, Vo M, Malik J. Long-term human motion prediction with scene context. In: Proc. ECCV 2020.
29. Yue S. Human motion tracking and positioning for augmented reality. J Real-Time Image Proc. 2021;18(2):357–68. https://doi.org/10.1007/s11554-020-01030-6.
30. Hernandez A, Gall J, Moreno F. Human motion prediction via spatio-temporal inpainting. In: 2019 IEEE/CVF International Conference on computer vision (ICCV), pp. 7133–7142. IEEE, Seoul, Korea (South) 2019. https://doi.org/10.1109/ICCV.2019.00723.
31. Amirian J, Hayet J-B, Pettre J. Social ways: learning multi-modal distributions of pedestrian trajectories with GANs. In: Proc. CVPR Workshops 2019.
32. Liang J, Jiang L, Niebles JC, Hauptmann AG, Fei-Fei L. Peeking Into the Future: Predicting Future Person Activities and Locations in Videos. In: Proc. CVPR 2019.
33. Harvey FG, Yurick M, Nowrouzezahrai D, Pal C. Robust motion in-betweening. ACM Trans Graph. 2020. https://doi.org/10.1145/3386569.3392480.
34. Mao X, Li Q, Xie H, Lau RYK, Wang Z, Smolley SP. Least squares generative adversarial networks. In: IEEE International Conference on computer vision (ICCV), 2017; pp. 2813–2821. https://doi.org/10.1109/ICCV.2017.304.
35. Lou H, Chai J. Example-Based Human Motion Denoising. IEEE Trans Vis Comput Graph. 2010;16(5):870–9. https://doi.org/10.1109/TVCG.2010.23. (**Conference Name: IEEE Transactions on Visualization and Computer Graphic**).
36. Piltaver R, Cvetković B, Kaluza B. Denoising human-motion trajectories captured with ultra-wideband real-time location system. Informatica. 2015;39:311–22.
37. Dagioglou M, Tsitos AC, Smarnakis A, Karkaletsis V. Smoothing of human movements recorded by a single RGB-D camera for robot demonstrations. In: The 14th PErvasive Technologies Related to Assistive Environments Conference. PETRA 2021, pp. 496–501. Association for Computing Machinery, New York, NY, USA 2021. https://doi.org/10.1145/3453892.3461627.
38. Memar Ardestani M, Yan H. Noise reduction in human motion-captured signals for computer animation based on B-spline filtering. Sensors. 2022;22(12):4629. https://doi.org/10.3390/s22124629. (**Number: 12 Publisher: Multidisciplinary Digital Publishing Institute**).
39. Kim SU, Jang H, Kim J. Human motion denoising using attention-based bidirectional recurrent neural network. In: SIGGRAPH Asia 2019 Posters. SA '19, pp. 1–2. Association for Computing

Machinery, New York, NY, USA 2019. https://doi.org/10.1145/3355056.3364577. Accessed 4 Aug 2022.
40. Cui Q, Sun H, Li Y, Kong Y. A deep bi-directional attention network for human motion recovery. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence. IJCAI'19, 2019; pp. 701–707. AAAI Press, Macao, China.
41. Lugaresi C, Tang J, Nash H, McClanahan C, Uboweja E, Hays M, Zhang F, Chang C-L, Yong MG, Lee J, Chang W-T, Hua W, Georg M, Grundmann M. MediaPipe: a framework for building perception pipelines. Technical Report arXiv (June 2019). https://doi.org/10.48550/arXiv.1906.08172. arXiv:1906.08172 [cs] type: article.
42. Yuan Y, Kitani K. DLow: diversifying latent flows for diverse human motion prediction. In: Vedaldi A, Bischof H, Brox T, Frahm J-M. editors. Proc. European Conference on Computer Vision - ECCV 2020, vol. 12354, pp. 346–364. Springer, Cham 2020. https://doi.org/10.1007/978-3-030-58545-7_20. Series Title: Lecture Notes in Computer Science.
43. Martínez-González A, Villamizar M, Odobez JM. Pose transformers (potr): Human motion prediction with non-autoregressive transformers. In: IEEE/CVF International Conference on computer vision—Workshops (ICCV), 2021.
44. Wandt B, Ackermann H, Rosenhahn B. 3D reconstruction of human motion from monocular image sequences. IEEE Trans Pattern Anal Mach Intell. 2016;38(8):1505–1516. https://doi.org/10.1109/TPAMI.2016.2553028.
45. Carnegie Mellon University: CMU MoCap Dataset. 2003. http://mocap.cs.cmu.edu. Accessed 28 July 2023.
46. Mahmood N, Ghorbani N, F. Troje N, Pons-Moll G, Black MJ. Amass: Archive of motion capture as surface shapes. In: The IEEE International Conference on computer vision (ICCV), 2019. https://amass.is.tue.mpg.de. Accessed 28 July 2023.
47. Ionescu C, Papava D, Olaru V, Sminchisescu C. Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Trans Pattern Anal Mach Intell. 2014. vol. 36, no. 7; pp 1325-1339. https://doi.org/10.1109/TPAMI.2013.248.
48. Ionescu C, Li F, Sminchisescu C. Latent structured models for human pose estimation. In: IEEE International Conference on computer vision (ICCV), 2011; pp. 2220–2227. https://doi.org/10.1109/ICCV.2011.6126500.
49. Gopinath D, Won J. Fairmotion—Tools to load, process and visualize motion capture data. Github 2020. https://github.com/facebookresearch/fairmotion. Accessed 16 May 2022.
50. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. Shortcut learning in deep neural networks. Nat Mach Intell. 2020;2(11):665–73. https://doi.org/10.1038/s42256-020-00257-z.