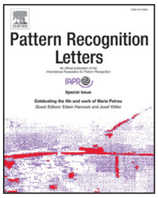
Pattern Recognition Letters 165 (2023) 138–145

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Sticks and STONES may build my bones: Deep learning reconstruction of limb rotations in stick figures☆

Francisco Fernandes[a,*], Ivo Roupa[b], Sérgio B. Gonçalves[b], Gonçalo Moita[c], Miguel Tavares da Silva[b], João Pereira[a,c], Joaquim Jorge[a,c], Richard R. Neptune[d], Daniel Simões Lopes[a,c]

[a] *Instituto de Engenharia de Sistemas e Computadores - Investigação e Desenvolvimento, Lisbon, Portugal*
[b] *Instituto de Engenharia Mecânica, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal*
[c] *Department of Computer Science and Engineering, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal*
[d] *Department of Mechanical Engineering, University of Texas, Austin, United States of America*

ARTICLE INFO

Article history:
Received 29 May 2022
Revised 10 November 2022
Accepted 19 December 2022
Available online 21 December 2022

Edited by Prof. S. Sarkar

Keywords:
Stick figure
Pose estimation
Longitudinal rotations
Machine learning
Neural networks

ABSTRACT

Monitoring and analyzing physical activity is becoming an important task in both clinical and non-clinical settings. To accomplish this desideratum, stick figures are often used as abstractions of human poses and movements by representing body segments as straight lines (sticks). Despite their straightforwardness, this minimalist representation is incomplete as it lacks the segments' longitudinal rotations, and therefore, is insufficient for applications requiring full 3D kinematic data. We introduce STONES, an advanced machine learning approach for estimating longitudinal body segment rotations of based on stick figures defined from a minimal set of body points. Our approach relies on a recurrent deep neural network, which takes 3D joint positions from a minimalist stick figure representation, such as those acquired by conventional depth camera sensors, and completes it with accurate longitudinal segment rotations. We validated our approach via a test scenario based on exergaming activities (e.g., lunges, squats, and kicks), which are becoming an emerging trend in several healthcare sectors, and our estimations show a fit above 98% and mean errors of approximately 1 °. Our deep learning approach effectively surpasses other machine learning-based strategies and closely matches the accuracy of state-of-the-art motion capture systems while running at real-time speeds.

1. Introduction

Kinematic analysis based on *stick figure* representations [1] plays an important role in the quantitative evaluation of human movement, being fundamental in areas such as biomechanics, sport science, clinical movement, and video games. As a minimalist abstraction, the stick figure represents the human form as a kinematic chain composed of body segments (i.e., line segments) defined by joints and endpoints. Stick figures require tracking and processing spatio-temporal information about body segments to perform kinematic analyses. This often results in an animated stick figure representation consisting of joint positions connecting rigid body segments. This representation can describe the key articulated body points and possible motion ranges, thus providing a succinct and accurate human pose and movement model.

Marker-based systems provide the best approach for precise and accurate motion capture of stick figures, as several cameras are used to track active/reflective markers, which are placed on the person's body, being considered the gold standard system in human motion analysis [2]. However, marker-based systems have several drawbacks such as high cost, long setup times, lack of portability, introduction of systemic experimental errors due to incorrect marker placement [3] or to unwanted marker displacements caused by soft tissue artifacts [4].

In contrast, markerless optical systems, such as single video, time-of-flight or stereoscopic cameras, resort to color and/or depth images that, once processed through advanced computer vision techniques, estimate joint positions in real-time. Systems, such as

☆ Editor: S. Sarkar.

* Corresponding author.

E-mail addresses: francisco.fernandes@ist.utl.pt (F. Fernandes), ivo.roupa@tecnico.ulisboa.pt (I. Roupa), sergio.goncalves@tecnico.ulisboa.pt (S.B. Gonçalves), miguelsilva@tecnico.ulisboa.pt (M.T. da Silva), jap@inesc-id.pt (J. Pereira), jaj@inesc-id.pt (J. Jorge), rneptune@mail.utexas.edu (R.R. Neptune), daniel.lopes@inesc-id.pt (D.S. Lopes).

https://doi.org/10.1016/j.patrec.2022.12.012
0167-8655/© 2022 Published by Elsevier B.V.

the *Kinect One* sensor [5], allow for easy and quick motion acquisition protocols, which provides lower costs and higher portability. However, these systems present several disadvantages, namely low accuracy in the estimation of the joint positions, occlusions of the segments with respect to the cameras and low acquisition frame rate. Some of these drawbacks can be handled by resorting to fault detection approaches and studying the influence of disturbances, modeling errors and other various uncertainties in these real systems [6–8].

Another significant shortcoming of these sensors is the inaccurate determination of segment orientations, in particular their longitudinal rotations [9,10]. This issue is a major drawback, since the longitudinal rotations of the upper and lower limbs play a key role in various human activities, such as grooming, feeding, dressing, bathing, toileting, walking, or running [11]. Moreover, the precise quantification of the longitudinal rotations is essential not only in the areas of the analysis of pathological human movement and physical rehabilitation [12,13] but also in the evaluation of sports performance [14], ergonomics [15], and digital animation [16].

To overcome these limitations and to augment motion capture data, this work proposes to use machine learning (ML) algorithms for reconstructing the missing longitudinal rotations of the body segments by using minimal marker sets that rely on the most basic information about stick figures. The proposed framework takes as input the 3D coordinates of each major body joint and extremity at each time frame (a total of 24 joints for common depth cameras), and outputs the corresponding longitudinal rotations of limb body segments (a total of 8 body segments: 2 forearms, 2 arms, 2 thighs, 2 legs).

While other studies have addressed similar Inverse Kinematics (IK) problems, our problem statement is different: previous work has considered extra data (image, video, point cloud, etc.) to reconstruct rotations, while we do not and show how the longitudinal rotations can be derived from a small sparse set of 3D points without any extra (image or video) data. The same way that a properly trained deep neural network is able to reconstruct, for instance, a heavily corrupted or noisy image even in the absence of the original color data [17], our work aims to develop an approach capable of learning how to reconstruct accurate longitudinal rotations from minimal data.

Our approach relies on supervised learning techniques to address this reconstruction problem. Specifically, we use a Recurrent Deep Neural Network (RNN) in an approach we call *STONES - Supervised Training for Orientable Neurally-Estimated body Segments*. To validate our *STONES* methodology, we compare the reconstructed rotations to the real data provided by a high-resolution marker-based system. Our results exhibit a close match between predictions and known rotations. Additionally, this culminated in a new motion capture database of physical exercises that feature high-quality orientation data, which is suitable for rehabilitation and other health related activities. Furthermore, our results were compared with and performed better than standard ML algorithms (OLS and SVR) and state-of-the-art video-based Deep Learning (DL) techniques, such as VIBE [18] and PARE [19].

2. Background

Methods for human pose estimation: Marker-based techniques were the first to be used for human pose estimation and have been evolving ever since [20]. One such example is *MoSh* [21], where both body shape and motion are estimated simultaneously using a motion-capture setup consisting of a sparse set of a few dozen markers. More recently, marker-less approaches have been developed for the reconstruction of human poses, based solely on non-intrusive video or depth data.

Depth sensors: Another powerful tool used in 3D human pose estimation are depth or stereoscopic sensors, with many different approaches able to extract 3D pose and shape from a single depth image, such as *DoubleFusion* [22] and *SimulCap* [23], which are even able to incorporate non-rigid deformations from clothing. Although widely used for skeletal tracking, the most frequently used depth-camera models lack the ability to provide limb orientations, or their joint angle estimations in general do not provide the required accuracy for many practical applications [24], and thus this data needs to be further post-processed by other methods to yield accurate segment orientations.

Deep learning: For human skeletal tracking and pose estimation from static or dynamic image data, convolutional neural network (CNN) based approaches have been the most popular and effective. This includes notable examples such as *DeepPose* [25] and *HRNet* [26] for the 2D pose from still images, or *A-NeRF* [27] and *DeepCap* [28] for 3D poses from single monocular RGB-camera videos. More advanced methods rely on attention mechanisms and use temporal information, such as VIBE [18] or PARE [19]. Many of these techniques are based on the 3D body model format SMPL [29], whose 3D mesh is parameterized by joint angles and a low-dimensional linear shape space.

Angles from minimal marker sets: VNect [30] and XNect [31] produce similar results to RGB-D cameras such as Kinect, but using a single RGB camera. They rely on an IK implementation where a CNN regresses 2D and 3D joint positions simultaneously, and produce temporally stable joint angles for a 3D skeleton. However, they require that both 2D and 3D predictions be combined to perform the kinematic skeleton fitting step containing the joint angles, which cannot be decoupled.

MotioNet [32] is a data-driven IK approach using a DNN that directly outputs a kinematic skeleton from a monocular video, where the network learns to infer 3D joint rotations directly from training data from real human motions. However, the input of this method are 2D positions, with the 3D positions only appearing at the end of the procedure, and therefore is not suitable for 3D marker data. In other work, Yiannakides et al. [33] use again a single monocular RGB camera but resort to searching on a large database of 2D multi-view joint projections to match 2D to 3D correspondences. But similarly, this model does not handle 3D data as input.

In Adult2Child [34], Dong et al. presented a method to translate adult motion into child-like motion from adult motion capture data and proposed an algorithm for transforming positional to rotational data by computing joint angles from joint positions. However, they noted the ambiguity in the reconstructed motion since the roll axis information, corresponding to the longitudinal rotations, cannot be fully recovered, and thus they specifically set it to zero.

Recently, Roupa et al. [35] also presented an efficient geometric method based on the motion envelopes feature to estimate the longitudinal rotations of stick figure models, which considered the shape of the surface traced by each model segment in space over time.However, this method presents several limitations, namely it cannot depict pure axial rotations or compute the initial angular position of the segments.

As previously described, many recent approaches in the literature focus on 3D pose estimation and reconstruction of movement based on still images, video, or depth data. However, although a few deal with minimal marker sets or joint rotation reconstruction, such studies consider other data sources (e.g., image, video, point cloud) or were not created to accept existing 3D joint coordinates as input. In addition, to our knowledge, no other work has addressed this problem with such sparse input data (i.e. from two 3D points per body segment, and with no other additional supporting data) and can reconstruct all six segment degrees of freedom.

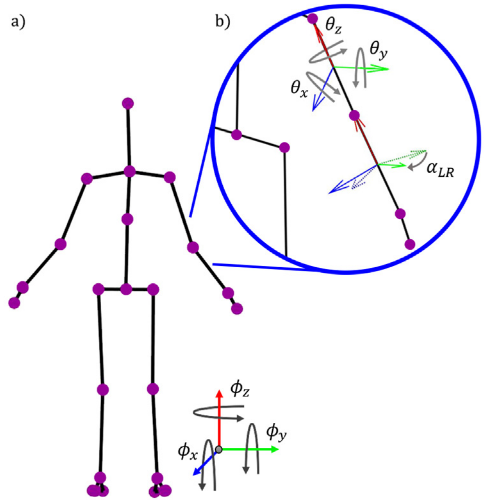


Fig. 1. (**a**) Representation of the biomechanical model composed of 20 segments (black lines) and 24 points (joints and extremity points - purple dots); **(b)** Detailed representation of the left upper arm and forearm segments. Solid vectors represent the local reference frame of each segment; dashed vectors depict the local reference frame of each segment orientated according to the projection of the medial-lateral vector of the parent body in the plane normal to the segment.

3. Data acquisition and representation

In this section we describe how data were acquired and represented in order to be processed by our approach.

3.1. Problem formulation

Under a rigid body assumption, stick figures with a minimal set of joints (i.e., that represent each body segment as a straight line connecting two contiguous joints) hold a challenging problem since, with only two non-overlapping points per rigid body segment, it is possible to determine only five degrees-of-freedom (three translations, two rotations). That is, as body segments consist of one-dimensional line objects in a three-dimensional space, the rotation around their own longitudinal axis (i.e., its *longitudinal rotation*, represented as θ_Z in Fig. 1(b)) cannot be determined by its two defining joint points alone. And thus, this presents an ill-posed problem that does not have a unique solution.

However, the ill-posedness of the regression problem can be circumvented by exploring ML techniques that take advantage of the remaining joints' positional data to infer these missing longitudinal rotations, for instance by internally exploiting correlations between segments, identifying indirect micro-displacements in adjacent joints, or detecting other compensatory movements in other body parts. The advantage of *deep neural networks* over other traditional ML approaches is that none of these concealed features needs to be manually identified or engineered since they can be automatically *learned* by the network with enough training data.

3.2. Acquisition of kinematic data

Experimental data acquisition for the training and testing of the neural networks consisted of two major steps. The first step included a method for the selection of movements to be acquired, while the second step considered the acquisition and processing of the kinematic data to be used as input to the neural network algorithm. This culminated in a database that we named the *LBL RollingSticks Database*.

A procedure to select the most relevant movements performed during fitness and gym workouts from an initial set of 100 different movements was performed by 10 fitness professionals. This identified ten different uniplanar movements, namely lunges, squats and their variants, in addition to two multiplanar movements, namely cross-jab-hook and cross-jab-kick, in order to assess the robustness of the neural network in less patterned movements. A detailed description of the selection methodology and the adopted criteria can be found in Supplementary Materials (SM1 and SM2).

The study was approved by the ethics committee of *Instituto Superior Técnico* in January 2020 (Ref. nr. 1/2020 (CE-IST)), consisting of a group of 16 volunteers with different levels of physical activity performing at least seven valid repetitions of each of the selected movements. The data was acquired using an optical MOCAP system composed by 14 Infrared ProReflex 1000 cameras (Qualisys©, Göteborg, Sweden) with an acquisition frequency of 100 Hz, resulting in a total of 761,234 frames to be used as inputs for the neural network.

A biomechanical model with 20 segments, seven extremities points (e.g., tip of the head, hands and feet) and 17 joints was developed in MATLAB (MathWorks©, Natick, USA) (see Fig. 2e), based on the Kinect One human model. A system of 70 reflective markers used for calibration and tracking was implemented to allow for the rigorous estimation of all the segments orientations (see Fig. 2a to c). A detailed description of the adopted marker set protocol and biomechanical model can be found in Supplementary Materials (SM3 and SM4).

3.3. Input and output

The goal of our approach was to predict the longitudinal rotations of the limbs segments (i.e., upper arms, forearms, thighs and legs for both sides) based solely on the location data from the body joints. Thus, the STONE model's input is the set of X, Y and Z floating point coordinates (in meters) of the 24 tracked body joints (seen in Fig. 2(e) as purple circles), represented as a vector $\vec{v}$ of size 72 for each tracked frame. The output consists of a vector $\vec{\hat{\theta}}$ of size eight corresponding to the floating point values for the Z component of the relative longitudinal rotation angles (in radians) of the stick figure's eight limb segments.

3.4. Data normalization

Before being input into the network, the input joints' spatial coordinates were normalized by translating and rotating all the points so that the coronal plane is parallel to the camera and its center is at the origin of the global reference frame. These preprocessing measures were useful to acquire a higher data robustness from people with different heights and body sizes, as well as to help the algorithm be insensitive to camera distance and orientation.

4. Network architecture

Our STONES approach relies on a Recurrent Deep Neural Network (RNN) to perform supervised regression by learning how to reconstruct angles from spatial coordinates. A graphical representation of its general layout can be found in Fig. 3. The RNN is able to handle a data sequence by processing multiple steps consecutively while maintaining information about the steps seen so far. Since motion-tracking data is a type of spatio-temporal data, an RNN was used to exploit the temporal aspect of multiple time steps.

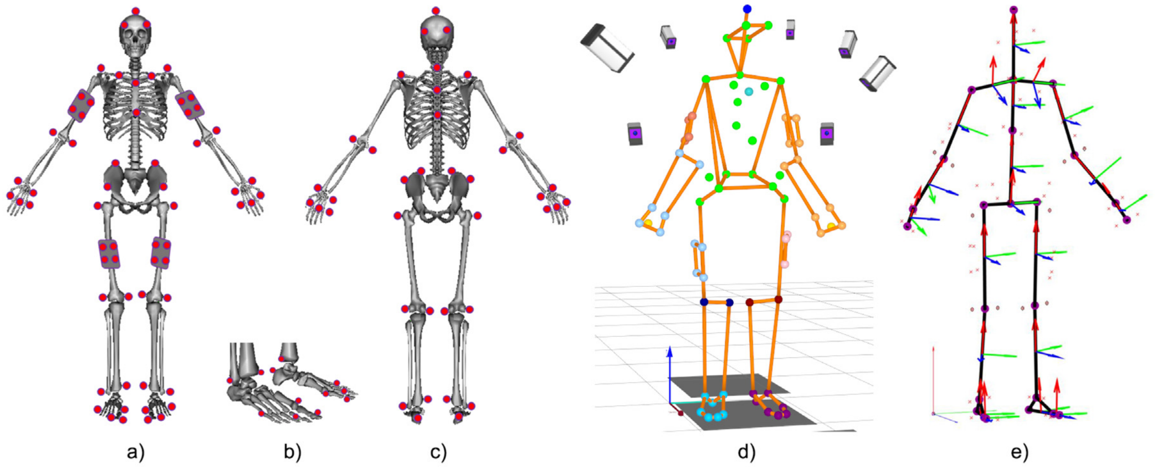


Fig. 2. Marker Set Protocol composed of 70 markers: **(a)** anterior view; **(b)** foot markers; **(c)** posterior view; **(d)** Experimental Acquisition Model; **(e)** Biomechanical Model.

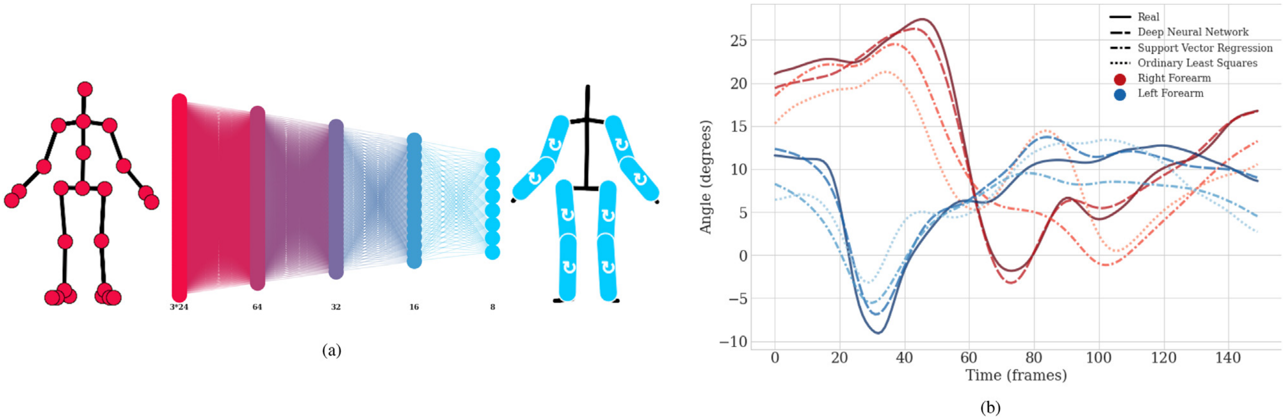


Fig. 3. **(a)** General layout of the RNN used in the STONES implementation, taking as input the 24 joint coordinates and estimating the rotations for the 8 body segments by going through several layers inside the neural network. **(b)** Known values and predictions of the left and right forearm rotations by the 3 different tested ML approaches across the first 150 frames of one of the captured actions.

4.1. Network parameters

We carried out a thorough exploration using cross-validation to determine an adequate network layout to our particular problem of performing regression of 8 angle values from multiple coordinate values. We chose the optimal STONES network attributes by hyper-parameter search over the number of hidden layers, neurons per hidden layer, activation functions, training optimizers and the number of frames to process simultaneously.

We arranged the layers in a funnel/pyramid layout, as depicted in Fig. 3, analogous to the *encoder* section of a regular *Autoencoder* DNN [36], with the search performed over several numbers and sizes of hidden layers. The final network layout was settled as a fully-connected recurrent neural network with five layers with sizes 72, 64, 32, 16, and 8, with a total of 12,792 wt variables.

We evaluated the length of the data sequence to input into the network, ranging from 2 to 10 time steps, and concluded that 2 frames were enough to accurately capture the limbs' rotation, with the increase in the number of input frames generating at most a 0.6% improvement in the scores. After evaluating 11 distinct activations and 8 optimizer alternatives, we selected the Softplus [37] activation function, while adopting AdaMax [38] as the optimizer algorithm to train the network, together with early-stopping.

4.2. Objective function and metrics

Since angle estimation corresponds to a numerical regression problem, the loss function used was a combined version of the *Mean Squared Error* (MSE), defined by:

$$MSE = \frac{1}{8}\sum_{i=1}^{8}\left(\frac{1}{n}\sum_{j=1}^{n}(\theta_j^i - \hat{\theta}_j^i)^2\right) \tag{1}$$

This meant that while training our network, for each rotation i and for each frame j, we aimed to minimize the squared difference between the true angle θ_j^i and our predicted angle $\hat{\theta}_j^i$, averaged over all n frames and the 8 rotations.

Another commonly used metric for regression problems is the *Coefficient of Determination*, which returns the proportion of the observed variation in the data that can be explained by the model, and in our case was computed as:

$$R^2 = \frac{1}{8}\sum_{i=1}^{8}\left(\frac{\sum_{j=1}^{n}(\hat{\theta}_j^i - \bar{\theta}^i)^2}{\sum_{j=1}^{n}(\theta_j^i - \bar{\theta}^i)^2}\right) \tag{2}$$

where $\bar{\theta}^i$ is the average of the ith rotation over all n true values. This provides a measure of how well our dependent variable (an-

Table 1
R^2, MSE and RMSE scores of the method's predictions, for each one of the eight individual rotations.

	Right				Left			
	Forearm	Arm	Leg	Thigh	Forearm	Arm	Leg	Thigh
R^2	0.979	0.997	0.957	0.997	0.981	0.997	0.968	0.998
MSE	2.32	0.32	1.38	0.33	2.22	0.34	1.63	0.25
RMSE	1.52	0.56	1.17	0.57	1.49	0.58	1.28	0.50
ROM	180	135	90	90	180	135	90	90
Captured	92.4	99.5	51.3	81.1	98.1	100.9	57.7	84.9
Estimated	92.4	99.5	47.9	81.1	97.8	98.8	57.3	84.2

gles) and its variability are being predicted by the independent variables (coordinates).

Due to the specific domain of our problem which revolves around limb orientations, in order to assess and compare our results, we also introduced a simple yet informative metric similar to the mean absolute error, the *Mean Per Limb Angle Error* (MPLAE) defined by:

$$MPLAE = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{8}\left|\frac{\hat{\theta}_j^i - \theta_j^i}{8}\right| \qquad (3)$$

which is a straightforward way to indicate the difference between the angle estimations and their true values for each of the 8 limbs across all considered test frames.

5. Results and discussion

In this section we present our results on accuracy for all eight limb segments, and compare the performance against other approaches.

5.1. Estimation accuracy

The predicted values for all eight rotation angles closely matched the true values for each of the limb segments considered (Table 1). The predictions show a close fit with the data variability between 95% and 99%, as measured by the R^2 score, with the thighs being the most accurate, followed by upper arms, forearms and legs. In terms of error values, the RMSE score oscillates around 1 °, again with the thighs presenting the best behaviour and the forearms being the most problematic, which is consistent with their respective anatomical range of motion (ROM) [39], but never exceeding 2 degrees.

5.2. Comparison with other ML methods

We compared our STONES solution against two standard ML techniques, the *Ordinary Least Squares* (OLS) method and *Support Vector Regression* (SVR) based on *libsvm* [40] and featuring a non-linear radial basis function kernel. Comparison of these three distinct approaches using different metrics are presented in Table 2 and Fig. 3. All methods has similarly low RMSE values below 5 degrees, but the RNN values were considerably lower (~1°). The three approaches are also clearly distinguishable when looking at R^2 scores. The linear OLS method was only able to predict ~76% of the data variability, while the non-linear methods achieved at least a 10% higher rate, in the case of SVR. The RNN yielded the best result with a fit close to 98%, which is 12% better than SVR.

The testing set processing speeds were also measured for each method and the times per sample (in milliseconds) are presented in Table 2. Although all methods are suitable for real time movement analysis, SVR is several orders of magnitude slower. The basic OLS approach is the fastest but shows the worst fit for the data.

Table 2
R^2, MSE and RMSE scores and frame processing times for each ML method.

		OLS	SVR	STONES
Evaluation subset	R^2	0.766	0.859	**0.984**
	MSE	19.315	11.082	**1.097**
	RMSE	4.395	3.329	**1.047**
	Time	**0.001**	2.135	0.033
New action data	R^2	0.725	0.702	**0.945**
	MSE	10.191	9.860	**1.983**
	RMSE	3.192	3.140	**1.408**
New subjects data	R^2	0.621	0.743	**0.866**
	MSE	37.944	25.173	**12.568**
	RMSE	6.160	5.017	**3.545**

The STONES approach is capable of processing more than 30,000 frames per second, which considering the additional computational overhead generating the stick figure visualization, is fast enough for practical applications.

5.3. Temporal coherence

We also studied the performance of all three methods from a continuous frame by frame perspective in order to examine the sequential evolution of each prediction. For this, we analyzed distinct fragments of different actions characterized by high variance in their rotations. Figure 3 details an example of such a segment, reporting a worst case scenario of the most problematic forearm rotations. All three approaches yield smooth approximations without discontinuities, although the RNN line more closely matches the true progression of both rotations, especially noticeable in the extremities of the considered frame interval. This is followed by SVR and then OLS, which features the less accurate tracking of the original shapes. The used MSE loss function inherently ensures the temporal smoothness of the results, since the motion of the input joint coordinates are also smooth across time.

5.4. Results per action

Different movement actions are associated with variable experimental ROMs for each segment, due to the different sets of limbs required to execute each particular action. By analysing the errors in each action (Fig. 4a), one can see that none of the captured actions is associated with a noteworthy higher or lower error value, although the rotations of both forearms seem to be particularly more challenging to estimate across all actions, closely followed by both legs. This is specially true in explosive or rapidly changing forearm movements such as both cross-jab actions. Additionally, analysing the error distribution with the variability of the rotations' amplitude inside each action and subject, depicted as a heat-map in Fig. 4b, shows that the fluctuation of the error spreads are uniformly throughout the different experimental ROMs, centered around the origin, and present no significant bias directed at shorter or wider ROMs.

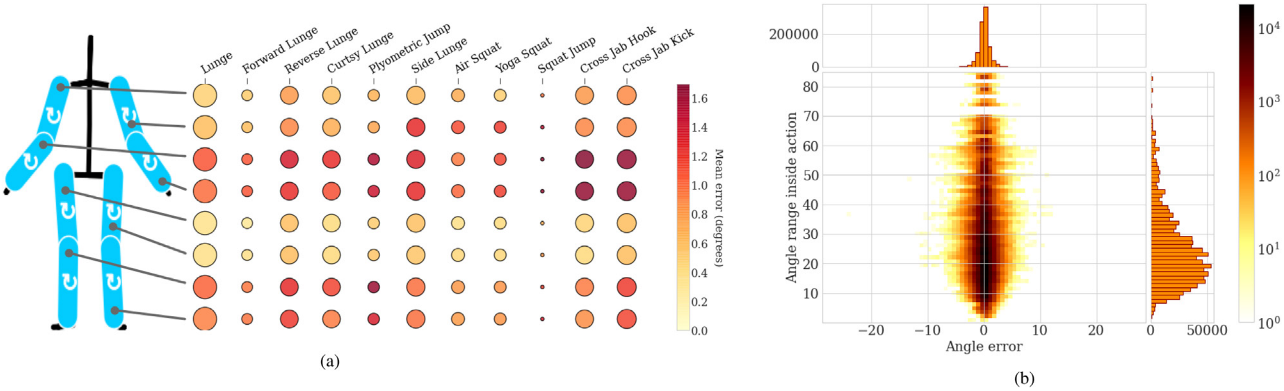


Fig. 4. **(a)** Average of all prediction errors for each action/rotation pair. Circle color matches error value while radius is proportional to the number of frames of each action in the evaluation subset. **(b)** Heat map of the errors of all estimated angles combined, according to the amplitude of the corresponding limb rotation in that particular action/subject event. The darker the color, the higher number of angle instances where that condition was observed (in logarithmic scale). The closer to the vertical line at zero, the better the performance.

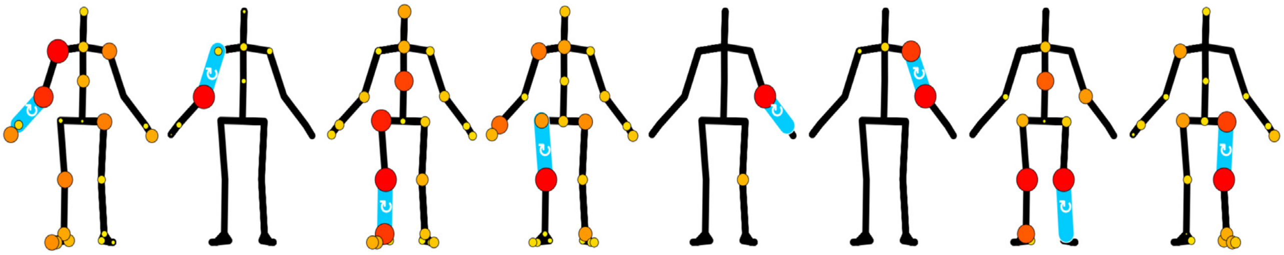


Fig. 5. Influence of each input joint in the prediction of each output rotation. The stronger the joint contribution, the larger its displayed radius.

5.5. Evaluation on new subjects and new action

To assess the ability of the approach to estimate the angles for any subject, we carried out an additional evaluation by testing the STONES method on two unfamiliar subjects and one unfamiliar action, both outside the population we used to train the model. The new data set spanned over a total of 73,811 frames for the new subjects and 27,740 frames for the extra action. Table 2 presents the evaluation results in the bottom rows. Results for the new action were only reduced slightly to ∼95%, while maintaining similar error profiles. The data fit for new subjects decreased to ∼87% as the error increased to ∼3 degrees. While these results are still suitable for most applications, they show that the model would benefit from handling more training subjects with more heterogeneous characteristics before being exposed to new test subjects. Nevertheless, the STONES approach outperforms the two competing ML methods by a comfortable margin in all three statistical scores for the new subjects.

5.6. Joint influence per angle

Since neural networks are usually modeled as *black boxes*, we put additional efforts into obtaining a better interpretation of the derived model and its inner workings by exploring how much individual junctions contributed to estimating each angle. To this end, we reset the network and retrained it multiple times for a fixed number of epochs, each time taking off (or 'dropping') one of the different 24 input joints. By examining the decrease in the score of the predicted angles relative to their original base score, we could infer an approximate measure about the omitted joint's importance inside the network to estimate each angle. Each joint's influence in each angle is illustrated in Fig. 5, where the circle radius is proportional to the observed score deterioration in each joint/angle combination. The higher this observed loss, the more critical a specific joint is to predicting an angle. For instance, arm and forearm rotations are heavily influenced by the elbow and shoulder joints, as expected, with the right side requiring more supportive joints from all around the body, possibly due to the broader range of movements attributable to the right-handedness of subjects.

5.7. Comparison with DL methods

We emphasize that existing DL methods which output limb rotations require video data as input, which in our scenario is not available as we rely on much more simplistic kinematic data composed of only the stick figure's 3D joints, and therefore a comparison on similar grounds cannot be fairly achieved. In fact, in such a reduced set of kinematic data, such as the sparse stick figure representation used in this work, standard IK methods will fail as information to rebuild the segments' internal rotations is missing in the input data, thus preventing the successful reconstruction of these degrees-of-freedom.

Nevertheless, we further compared STONES against two other state-of-the-art video-based methods, VIBE [18] and PARE [19]. These two methods were run on existing video footage and their results are presented in Table 3. In order to analyze longitudinal rotations from the output, joint rotations were extracted from the SMPL format, converted to Euler angles taking into account the differences between T-pose and human reference position, upsampled using bi-cubic interpolation to account for the different frame rates, centered and synchronized with the ground-truth signal using cross-correlation to find the best frame lag due to different time duration, and finally their RMSE and MPLAE scores calculated. Results show that STONES always outperforms both approaches in terms of RMSE by at least ∼2 degrees in most cases, with accuracy differences reaching 15 degrees in both arms. When considering the MPLAE score, STONES results display similar be-

Table 3
MPLAE and RMSE scores of all tested DL methods for each one of the eight individual rotations.

	MPLAE	RMSE							
		Right				Left			
		Forearm	Arm	Leg	Thigh	Forearm	Arm	Leg	Thigh
STONES	**0.661**	**0.89**	**0.51**	**1.17**	**0.61**	**2.70**	**0.54**	**0.91**	**0.39**
VIBE	5.511	2.08	16.55	3.33	3.66	4.10	14.85	3.57	3.24
PARE	5.046	2.61	15.16	3.57	1.98	3.72	14.51	3.87	2.04

haviour, with errors consistently inferior to the other two approaches, on the order of 5 degrees.

6. Conclusion

We have developed a new motion processing approach called STONES to predict the longitudinal rotation angles of body segments represented by a stick figure. This abstraction consists of a minimal marker set representation of the human pose, constructed from merely kinematic data of a few body joints alone, namely the trajectories of 2 joints per body segment. STONES showcases high agreement with known rotations, fitting more than 95% of the data variability for each of the eight considered rotations and across all actions. Forearms exhibit a higher degree of error, while thigh predictions were the most accurate, which corresponds with their anatomical ROMs. Nevertheless, combined mean errors fell within $\sim$2 degrees.

The approach proved robust to a newly seen action and new subjects where the fit was reduced to 85% while still featuring a mean error of $\sim$3 degrees. These results indicate that the training dataset should include more subjects and movements in the future. Further tests revealed that our technique outperforms SVMs and two other state-of-the-art DL methods by estimating variability and error values using a minimal fraction of their kinematic data. Moreover, its predictions are temporally coherent, and its performance is suitable for real-time human kinematic applications.

This study demonstrated that our method could accurately estimate the stick figure's longitudinal rotations, representing anatomically correct rotations performed by segments of the human body. This method can be used to reliably augment and enrich existing skeletal tracking methods with orientation information for those that lack such data. In summary, STONES enables computationally inexpensive, fast and accurate human pose estimation for physical exercises without the disadvantages of more complex marker-based methods.

The assembled motion capture dataset (*RollingSticks*), as well as the STONES trained network model including the full source code to generate and run the algorithm, is available at https://lbl.tecnico.ulisboa.pt/StreackerDB.html.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The work reported in this article was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with references UIDB/50021/2020, UIDB/50022/2020 (through IDMEC under LAETA), 02/SAICT/2017 and projects UTAP-EXPL/CA/0065/2017 and PTDC/CCI-COM/30274/2017.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patrec.2022.12.012.

References

[1] E.-J. Marey, La méthode graphique dans les sciences expérimentales et particulièrement en physiologie et en médecine, G. Masson, 1878.
[2] G. Saggio, F. Tombolini, A. Ruggiero, Technology-based complex motor tasks assessment: a 6-DOF inertial-based system versus a gold-standard optoelectronic-based one, IEEE Sens. J. 21 (2) (2021) 1616–1624.
[3] U. Della Croce, A. Leardini, L. Chiari, A. Cappozzo, Human movement analysis using stereophotogrammetry: Part 4: assessment of anatomical landmark misplacement and its effects on joint kinematics, Gait Posture 21 (2) (2005) 226–237.
[4] I. Roupa, M.R. da Silva, F. Marques, S.B. Gonçalves, P. Flores, M.T. da Silva, On the modeling of biomechanical systems for human movement analysis: a narrative review, Arch. Comput. Methods Eng. (2022) 1–44.
[5] Microsoft, Kinect for windows, 2014, (https://developer.microsoft.com/en-us/windows/kinect/), Accessed: 2020-05-20.
[6] Z. Xu, X. Li, V. Stojanovic, Exponential stability of nonlinear state-dependent delayed impulsive systems with applications, Nonlinear Anal. Hybrid Syst. 42 (2021) 101088.
[7] X. Song, P. Sun, S. Song, V. Stojanovic, Event-driven NN adaptive fixed-time control for nonlinear systems with guaranteed performance, J. Franklin Inst. (2022).
[8] Z. Zhuang, H. Tao, Y. Chen, V. Stojanovic, W. Paszke, Iterative learning control for repetitive tasks with randomly varying trial lengths using successive projection, Int. J. Adapt. Control Signal Process. 36 (5) (2022) 1196–1215.
[9] R.A. Clark, Y.-H. Pua, K. Fortin, C. Ritchie, K.E. Webster, L. Denehy, A.L. Bryant, Validity of the microsoft kinect for assessment of postural control, Gait Posture 36 (3) (2012) 372–377.
[10] M. Huber, A.L. Seitz, M. Leeser, D. Sternad, Validity and reliability of kinect skeleton for measuring shoulder joint angles: a feasibility study, Physiotherapy 101 (4) (2015) 389–393.
[11] S.M.B. Gonçalves, S.B.C. Lama, M.T. da Silva, Three decades of gait index development: acomparative review of clinical and research gait indices, Clin. Biomech. (2022) 105682.
[12] D.S. Lopes, A. Faria, A. Barriga, S. Caneira, F. Baptista, C. Matos, A.F. Neves, L. Prates, A.M. Pereira, H. Nicolau, Visual biofeedback for upper limb compensatory movements: a preliminary study next to rehabilitation professionals, EuroVis 2019 - Posters, The Eurographics Association, 2019. pp. –
[13] T. Alves, H. Carvalho, D. S. Lopes, Winning compensations: adaptable gaming approach for upper limb rehabilitation sessions based on compensatory movements, J. Biomed. Inform. 108 (2020) 103501.
[14] N.A. Trasolini, K.F. Nicholson, J. Mylott, G.S. Bullock, T.C. Hulburt, B.R. Waterman, Biomechanical analysis of the throwing athlete and its impact on return to sport, Arthroscopy Sports Med. Rehabil. 4 (1) (2022) e83–e91.
[15] A.C. McDonald, D.M. Mulla, P.J. Keir, Muscular and kinematic adaptations to fatiguing repetitive upper extremity work, Appl. Ergon. 75 (2019) 250–256.
[16] A. Aristidou, J. Lasenby, Y. Chrysanthou, A. Shamir, Inverse kinematics techniques in computer graphics: asurvey, Comput. Graphics Forum 37 (6) (2018) 35–58.
[17] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, T. Aila, Noise2noise: learning image restoration without clean data, in: International Conference on Machine Learning, 2018, pp. 2965–2974.
[18] M. Kocabas, N. Athanasiou, M.J. Black, VIBE: video inference for human body pose and shape estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5253–5263.
[19] M. Kocabas, C.-H.P. Huang, O. Hilliges, M.J. Black, PARE: part attention regressor for 3D human body estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11127–11137.
[20] T.L. Munea, Y.Z. Jembre, H.T. Weldegebriel, L. Chen, C. Huang, C. Yang, The progress of human pose estimation: a survey and taxonomy of models applied in 2D human pose estimation, IEEE Access 8 (2020) 133330–133348.

[21] M. Loper, N. Mahmood, M.J. Black, MoSh: motion and shape capture from sparse markers, ACM Trans. Graphics (TOG) 33 (6) (2014) 1–13.

[22] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, Y. Liu, DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7287–7296.

[23] T. Yu, Z. Zheng, Y. Zhong, J. Zhao, Q. Dai, G. Pons-Moll, Y. Liu, SimulCap: single-view human performance capture with cloth simulation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 5499–5509.

[24] T.M. Guess, S. Razu, A. Jahandar, M. Skubic, Z. Huo, Comparison of 3D joint angles measured with the kinect 2.0 skeletal tracker versus a marker-based motion capture system, J. Appl. Biomech. 33 (2) (2017) 176–181.

[25] A. Toshev, C. Szegedy, DeepPose: human pose estimation via deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1653–1660.

[26] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693–5703.

[27] S.-Y. Su, F. Yu, M. Zollhoefer, H. Rhodin, A-NeRF: surface-free human 3D pose refinement via neural rendering, arXiv preprint arXiv:2102.06199(2021).

[28] M. Habermann, W. Xu, M. Zollhofer, G. Pons-Moll, C. Theobalt, DeepCap: monocular human performance capture using weak supervision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5052–5063.

[29] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, M.J. Black, SMPL: a skinned multi-person linear model, ACM Trans. Graphics (TOG) 34 (6) (2015) 1–16.

[30] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt, VNect: real-time 3D human pose estimation with a single RGB camera, ACM Trans. Graphics (TOG) 36 (4) (2017) 1–14.

[31] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, M. Elgharib, P. Fua, H.-P. Seidel, H. Rhodin, G. Pons-Moll, C. Theobalt, XNect: real-time multi-person 3D motion capture with a single RGB camera, ACM Trans. Graphics (TOG) 39 (4) (2020). 82–1

[32] M. Shi, K. Aberman, A. Aristidou, T. Komura, D. Lischinski, D. Cohen-Or, B. Chen, MotioNet: 3D human motion reconstruction from monocular video with skeleton consistency, ACM Trans. Graphics (TOG) 40 (1) (2020) 1–15.

[33] A. Yiannakides, A. Aristidou, Y. Chrysanthou, Real-time 3D human pose and motion reconstruction from monocular RGB videos, Comput. Animat. Virtual Worlds 30 (3–4) (2019) e1887.

[34] Y. Dong, A. Aloba, S. Paryani, L. Anthony, N. Rana, E. Jain, Adult2child: dynamic scaling laws to create child-like motion, in: Proceedings of the Tenth International Conference on Motion in Games, 2017, pp. 1–10.

[35] I.F. Roupa, S.B. Gonçalves, M.T.d. Silva, R.R. Neptune, D.S. Lopes, Motion envelopes: unfolding longitudinal rotation data from walking stick-figures, Comput. Methods Biomech. Biomed. Eng. (2021) 1–12.

[36] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[37] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.

[38] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings, 2015. pp. –

[39] M. Clark, S. Lucett, et al., NASM Essentials of Corrective Exercise Training, Lippincott Williams & Wilkins, 2010.

[40] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst.Technol. (TIST) 2 (3) (2011) 1–27.