Weighted Average of Human Motion Sequences for Improving Rehabilitation Assessment

Ali Ismail-Fawaz[1], Maxime Devanne[1], Stefano Berretti[2], Jonathan Weber[1], and Germain Forestier[1,3]

[1] IRIMAS, Université de Haute-Alsace, France
{ali-el-hadi.ismail-fawaz,maxime.devanne,jonathan.weber,germain.forestier}@uha.fr
[2] MICC, University of Florence, Italy stefano.berretti@unifi.it
[3] DSAI, Monash University, Australia
germain.forestier@monash.edu

Abstract. While human motion analysis has been widely addressed in recent years, the specific task of rehabilitation motion assessment remains challenging due to the lack of available annotated data. To overcome this challenge, data augmentation can be considered. However, classical augmentation techniques applied to human motion sequences often result in meaningless movements. Moreover, in rehabilitation assessment, labels are often continuous values illustrating the quality of a movement. Hence, associating a continuous label to augmented data is not straightforward. In this work, we propose to address data augmentation using an averaging method, called *shapeDBA*, adapted to rehabilitation motion sequences represented as multivariate time series. We extend the original proposal by weighting the average, hence allowing us to infer continuous labels associated to augmented motion sequences. We evaluated our proposed method on the Kimore dataset. Experimental results show that our method generates coherent rehabilitation sequences that can be efficiently used to extend a small dataset for rehabilitation assessment.

Keywords: Rehabilitation assessment · Data extension · Synthetic motion data

1 Introduction

Human motion data is everywhere in our daily lives and has important applications in the medical sector such as in human rehabilitation assessment. It helps doctors and therapists to track patient progress, design personalized recovery plans and improve treatment outcomes. To capture human movement, while expensive and accurate Motion Capture (MoCap) systems are mainly used in the entertainment domain, they may be unsuitable for rehabilitation as they require the wearing of sensors or dedicated suits. Conversely, skeleton data extracted from videos are much less intrusive, and thus are increasingly adopted in rehabilitation centers.

In rehabilitation assessment, a challenging task involves analysing a rehabilitation exercise to predict a performance score corresponding to how well it

has been performed by a patient. Performance scores are continuous values often ranging from 0 to 100, hence estimating such scores can be modeled as an extrinsic regression problem. Extrinsic regression refers to predicting an output continuous variable based on input data samples using machine learning models [8].

Deep learning approaches have been successfully employed for various skeleton-based human motion analysis tasks, where consequent datasets are available. However, such approaches may be limited when a small amount of motion sequences is available, like in the medical field. This is the case for rehabilitation assessment, where acquiring real motion sequences performed by patients during their rehabilitation program is not straightforward [1]. The data collection task is further complicated when it comes to annotations, as associating a performance score to each acquired rehabilitation sequence is time consuming and requires the support of clinical experts. These reasons certainly explain the few amount of publicly available rehabilitation datasets as well as their limited size.

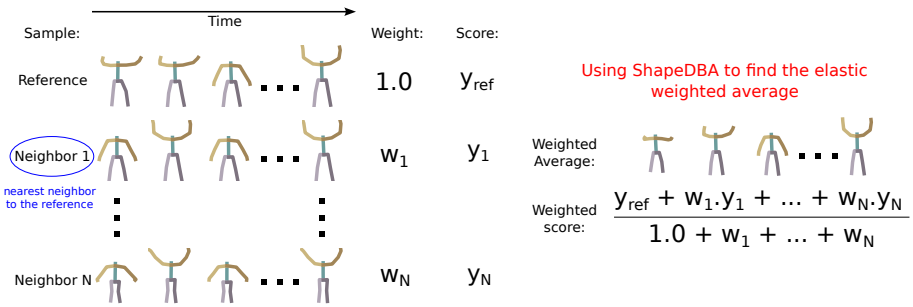

Fig. 1. Identifying N nearest neighbors for each reference in the dataset using Dynamic Time Warping (DTW). A weight is assigned to each neighbor based on the DTW measure to the reference. Using these weights, the weighted average sequence is computed via ShapeDBA, followed by calculating the weighted score to extend the regression training dataset with this new sample.

A common approach to address the lack of data in machine learning is data augmentation (DA), where new synthetic examples are created to increase the variability of the training data, improving the model's generalization capabilities. In the case of human rehabilitation assessment datasets, the most used method for generating new samples involves adding noise to existing real samples. While this technique can be effective in some cases [35], it may not always be the most appropriate approach, as it can lead to unrealistic and meaningless sequences.

Another way of generating synthetic sequences is using deep generative models, which have proven effective in other human motion tasks by generating realistic sequences [29]. However, these approaches present a paradox: if we do not have enough data to train a deep supervised model, how can we train a deep

generative model on such a small dataset? To address this issue, researchers in the time series domain often rely on averaging techniques [28] aiming at creating average representative sequences from a set of training samples [18].

In this work, we propose to study the use of a recent averaging method, ShapeDBA [17], on skeleton-based rehabilitation sequences. This approach has been previously used to generate synthetic time series datasets through a weighted elastic averaging method called Dynamic Time Warping Barycenter Averaging (DBA). More recently, Ismail-Fawaz et al. [17] introduced a novel averaging method, ShapeDBA, which adapts DBA to more accurately reflect changes in sequence shapes over time. In this study, we utilize the weighted version of ShapeDBA approach to generate various synthetic average sequences. As such averaging method is time consuming, it cannot be computed at each epoch of the training phase, as usually done in data augmentation. Differently, we generate several average sequences to extend the size of the original training dataset, before model's training. We refer to this as data extension (DE) to differentiate it from data augmentation.

In this work, we target the task of rehabilitation assessment, which is often formulated as an extrinsic regression problem where the goal is to predict a continuous performance score associated to a rehabilitation sequence. In the state-of-the-art, most of the work have considered DA or DE for classification purposes, where each synthetic sample is associated with a discrete label corresponding to the original sample that has been augmented. However, associating a continuous label to synthetic data is not straightforward. We hence propose to leverage the weighting strategy by inferring a weighted continuous label computed from true labels of a given set of considered samples. Our proposed setup is detailed in Figure 1.

It results in an extended dataset with both synthetic sequences and synthetic continuous labels, allowing deep learning models to generalize better for the task of rehabilitation assessment. The main contributions of this work are:

- We investigate the use of ShapeDBA for human skeleton sequences to generate average rehabilitation movements;
- We extend the averaging method by a weighting strategy that allows generating more various but yet coherent sequences;
- We leverage the weighting strategy for associating meaningful continuous labels to synthetic rehabilitation sequences;
- We consider the proposed method for extending a small real-world rehabilitation dataset and evaluate it on an extrinsic regression task.

2 Related Work

Skeleton-based approaches have shown promise for human motion analysis. Instead of considering the whole video stream, these approaches rely on humanoid skeleton data extracted from each frame of the video, and allow capturing human posture and movement. This skeleton modality has also been explored for

rehabilitation assessment in combination with classical Machine Learning algorithms such as Hidden Markov Models (HMM) [2] or Gaussian Mixture Models (GMM) [6]. More recently, Deep Learning approaches have been increasingly employed for rehabilitation assessment. One of the first Deep Learning methods applied to rehabilitation assessment from 3D skeleton data was proposed in [22], where a Long Short-Term Memory Network (LSTM) was considered. Inspired by the success of Convolutional Neural Network for time series classification [12,16,19], Ismayilzada et al. [20] investigated the use of various convolution-based architectures for rehabilitation assessment. Differently, a graph has been considered in [5] to represent the spatial and temporal relationship between human joints. Then, a Graph Convolutional Network (GCN) has been proposed for modeling and analysing the whole rehabilitation sequence. This promising method has been further extended in [24] by combining both positional and angular skeleton features with an attention fusion mechanism.

However, similarly to various fields, these Deep Learning models may suffer from overfitting and tend to not generalize well on new patients not seen during training [20]. This is particularly due to the lack of publicly available datasets for rehabilitation assessment. Moreover, while few datasets have been made available such as Kimore [3], UI-PRMD [32] or KERAAL [27], all of them include a limited number of rehabilitation sequences, thus complicating the learning task by deep models.

A well explored way of addressing data limitation and improving generalization of deep models is known as data augmentation (DA). DA involves introducing different variations in the training set at each epoch of the training phase, thus preventing the deep model to overfit the original training data. In 3D skeleton data, the most simple variations include geometric transformations [4] such as translation, rotation and scaling. However, such transformations do not affect the temporal dynamics of human motion, thus do not allow a deep model to learn various ways of performing rehabilitation exercises. To overcome this, many works consider applying random noise on skeleton sequences [4, 31]. Differently, as human motion is considered as a spatio-temporal data, temporal DA [30] can also be considered, such as shifting [21] or warping [11]. While noisy or temporal augmentations have been shown to be effective for skeleton-based action recognition [35], they may result in incoherent motion as they do not consider the data distribution.

Conversely, generative approaches aim at creating synthetic data from a set of training samples. In recent years, many deep generative models have been proposed for human motion [14, 29]. However, for rehabilitation assessment, such deep generative models may suffer from the limited amount of training data to correctly learn the data distribution and generate coherent rehabilitation sequences. In this case, other generative methods computing averages from a set of samples seem to be more appropriate. For temporal sequences, such as rehabilitation exercises, an elastic averaging approach is required to consider temporal alignment. Hence, the DTW Barycenter Averaging (DBA) has been proposed in [28] by leveraging Dynamic Time Warping (DTW) [25] for temporal align-

ment. Similarly, a more recent averaging method called ShapeDBA [17] proposed to leverage the ShapeDTW measure [36], which is more robust to noise. However, such averaging methods are often computationally costly and thus cannot be computed at each epoch of the training phase. Instead, synthetic averages can be computed only once to increase the size of the original training set prior to the training phase. To differentiate this strategy from DA, we refer to it as data extension (DE). In this work, we consider skeleton DE and propose to use the recent ShapeDBA averaging method in a weighted version to generate meaningful synthetic rehabilitation sequences.

3 Proposed Approach

3.1 Skeleton Sequence Representation

In this work, we are interested in assessing physical rehabilitation exercises represented by 3D skeleton sequences extracted from RGB-D videos. For each frame t of a sequence, the 3D position of each joint j of the skeleton is represented by three coordinates $x_j(t)$, $y_j(t)$ and $z_j(t)$. Let N_j be the number of joints the skeleton is composed of, the skeleton pose $p(t)$ at frame t is then represented by a $3N_j$ dimensional vector:

$$p(t) = [x_1(t), y_1(t), z_1(t), \ldots, x_{N_j}(t), y_{N_j}(t), z_{N_j}(t)]. \tag{1}$$

The whole skeleton sequence S of length T is represented by a multivariate time series of shape $(3N_j, T)$:

$$S = [p(1), p(2), \ldots, p(T)]. \tag{2}$$

Such a multivariate time series S represents the evolution of a body pose p through time.

Each sequence is associated with a continuous label y representing a performance score, i.e., how much the rehabilitation exercise is correctly performed.

3.2 Sequence Averaging using ShapeDBA

Sequence averaging is a task that is more commonly referred to as *prototyping* with the goal of finding a representative sample for a set of sequences. Given the nature of temporal ordering in sequences, especially human motion sequences, Petitjean et al. [28] argued the need of an elastic averaging approach to overcome the issues of traditional averaging techniques. The naive way of averaging temporal sequences is with the arithmetic mean; however, it assumes a perfect alignment between all samples, which is not the case in most datasets. Petitjean et al. [28] proposed the usage of elastic measures between sequences, such as Dynamic Time Warping (DTW) [25], and proposed DTW Barycenter Averaging (DBA). DTW is a similarity measure between two sequences that aligns two input sequences through a warping path, detecting some temporal distortions along the way, and performing the Euclidean Distance (ED) over the aligned series. DBA utilizes this algorithm through the following steps:

- **step 1:** Initialize the average sequence as one of the samples in the given dataset;
- **step 2:** Find the warping path between the average sequence and all other samples in the dataset;
- **step 3:** For each time stamp in the average sequence, replace its value with the barycenter of all aligned points from other samples in the dataset. The barycenter in this case refers to the arithmetic mean of all aligned values. Repeat **step 2** until convergence, meaning the average sequence no longer changes compared to a given threshold.

More recently, Ismail-Fawaz et al. in [17] argued that the usage of DTW in DBA can produce some out-of-distribution points in the output average sequence. This is due to the fact that aligning point-to-point between two time series presents some issues as argued in [36], for which they proposed ShapeDTW as a solution. ShapeDTW is a variant of DTW that aligns neighborhoods instead of points between two samples. Ismail-Fawaz et al. [17] hence proposed ShapeDBA, the DBA algorithm utilizing the ShapeDTW alignment method.

In this work, we utilize the ShapeDBA approach of averaging in order to create new samples and extend the training dataset.

3.3 Weighted Averages

As the manifold of rehabilitation motion sequences may be sparse, computing an average using the original shapeDBA method may result in a meaningless or incoherent average sequence. In order to follow the distribution of the motion sequences, a weighted average is more appropriate [7]. For computing a weighted average motion sequence, we employ the ShapeDBA algorithm and follow a similar strategy as in [7, 15]. For a given reference motion sequence, we generate a synthetic version by considering a neighborhood of n motion sequences. A weight of 1 is associated to the reference sequence S_{ref}, while for neighboring sequences S_i, the weight depends on the similarity with the reference and is computed as:

$$w_i = e^{\ln(0.5) * \frac{DTW(S_i, S_{ref})}{d_{NN}}}, \tag{3}$$

where d_{NN} corresponds to the DTW distance between S_{ref} and its nearest neighbor. This allows to give more impact on nearest sequences, while computing the weighted average sequences. Figure 2 illustrates the importance of a weighted average when the considered manifold is sparse, and the distribution of motion sequences is not spherical.

For each reference sequence S_{ref}, the computation of the weighted ShapeDBA results in a corresponding synthetic sequence $\hat{S}$. In order to associate a continuous label (score) to the synthetic data, the same weights are used but are first normalized using a min-max normalization. This ensures the weighted continuous labels keep in the original range between 0 and 1. Hence, the associated

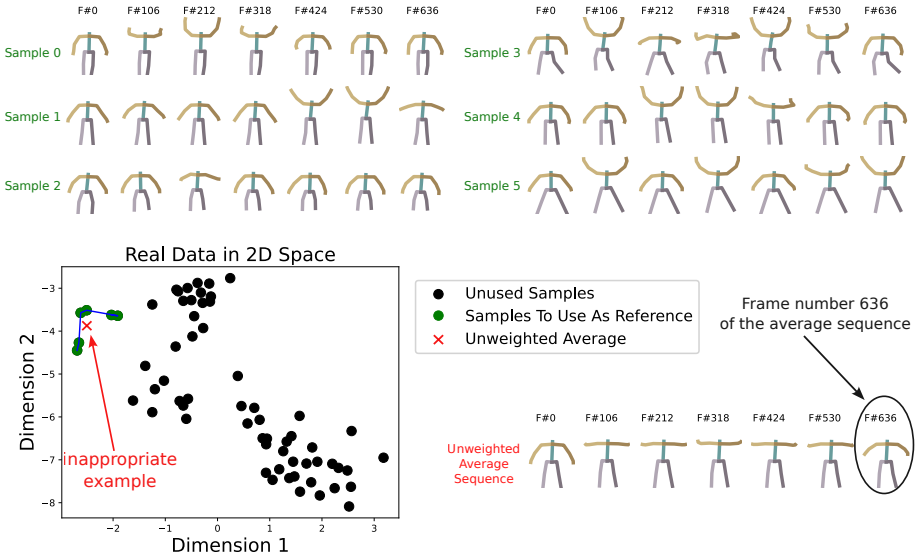

Fig. 2. Averaging six human motion sequences (from Example 0 to Example 5 on the top left/right) assuming a uniformly distributed weight produces an inappropriate example (bottom right skeleton sequence). Following the 2D t-SNE projection of these sequences (bottom left), the placement of the average sequence (bottom left in red) does not belong to the manifold of the original sequences (bottom left in green).

label $\hat{y}$ to a synthetic sequence $\hat{S}$ is computed as:

$$\hat{y} = \sum_{i=1}^{N+1} \bar{w}_i * y_i, \tag{4}$$

where $\bar{w}_i$ is the normalized weight associated to the i-th sequence in the considered set of size $N+1$ (the reference sequence and the N neighboring sequences).

4 Experimental Evaluation

We evaluated our proposed approach on two different aspects. The first one assesses the coherence of synthetic sequences, both qualitatively and quantitatively. The second investigates how synthetic sequences generated using our approach can be employed to help a deep learning model achieving extrinsic regression, i.e., predicting a continuous score associated with a rehabilitation sequence. The code is publicly available here: https://github.com/MSD-IRIMAS/Weighted-ShapeDBA-4-Rehab

4.1 Experimental Setup

Dataset: We used the Kimore dataset [3] including RGB-D videos collected by a Kinect sensor from which 3D skeleton sequences have been extracted, for five distinct rehabilitation exercises. Each sequence is associated with an evaluation score provided by clinical experts, ranging from 0 to 100 (normalized to the range $[0, 1]$). In this work, we discarded rehabilitation sequences with missing skeleton values. It resulted in 71 sequences per exercise corresponding to 40 healthy subjects and 31 unhealthy patients. We resampled all the sequences to the average length of 748 frames using the Fourier method in the *scipy* [33]. We employed a 5-fold cross-validation protocol but differently from what is usually done in the literature, we slightly adapted the 5-fold split in order to consider unhealthy subjects at inference. We believed this is more close to a real-world scenario. As a result, only sequences corresponding to unhealthy subjects were split into 5 folds. For each run, all sequences corresponding to healthy subjects and 4 unhealthy folds were used for training, while the remaining unhealthy fold was used for test. We utilize *aeon* [23], a python toolkit for time series machine learning, for the similarity measures and ShapeDBA.

Comparative Sets of Rehabilitation Sequences: Our comparative study includes different sets of rehabilitation sequences. The *reference set*, of size D corresponds to original sequences from the Kimore dataset. As baseline, we created a noisy version of each reference sequence by adding a random noise $\mathcal{N}(0, 0.1)$ to all skeletons of the sequence. We referred to this set of size D as the *noisy set*. We then created 5 additional sets using the proposed method generating synthetic sequences. For each reference sequence from the *reference set*, we generated a synthetic version using weighted ShapeDBA, detailed in Section 3.3, by considering 5 different neighborhoods of size $N = 1$ to $N = 5$. It resulted in 5 new sets of size D, denoted as *ShapeDBA NN1*, *ShapeDBA NN2*, *ShapeDBA NN3*, *ShapeDBA NN4* and *ShapeDBA NN5*.

4.2 Evaluation of Synthetic Data Coherence

When it comes to generative models, it is mostly difficult to utilize one way of evaluation. For human motion data, one approach is to assess how realistic generated sequences are visually. However, it is not enough to use human visualisation as argued in [26], and metrics are needed to quantify how reliable generated samples are. In this section, we propose both ways to assess this reliability: visually and numerically.

Generation Metrics

The task of extending a dataset, even through non-deep learning methods, is still considered as a generation pipeline. For this reason, we found it crucial to assess the fidelity and diversity [26] of generated samples. Fidelity and diversity are both aspects of quantitative evaluation of generative models. On the one

hand, fidelity metrics evaluate how close the real and generated samples are. The closer they are the more *reliable* the generated samples are for real world applications in terms of fidelity. On the other hand, diversity metrics evaluate how far samples are from each other, independently for real and generated samples. Moreover, the goal of a generative model is as well, most of the times, to produce a generated space as diverse as the real one.

In this work, we relied on two common metrics to assess fidelity and diversity: The Fréchet Inception Distance (FID) and the Average Pair Distance (APD).

Fréchet Inception Distance (FID). The FID metric [10] assesses how close real and generated samples are in terms of distribution. For computing it, the real and generated samples first go through a latent feature extraction step using a pre-trained deep learning model $\mathcal{F}$ (in our case trained on a regression task). Second, the mean and covariance are calculated for both feature space. Third, FID computes the distance between both distributions assuming the mean and covariance are of a Gaussian distribution. The mathematical formulation of FID is defined below:

$$FID(\mathcal{P}_1, \mathcal{P}_2)^2 = trace(\Sigma_1 + \Sigma_2 - 2(\Sigma_1.\Sigma_2)^{1/2}) + \sum_{i=1}^{f}(\mu_{1,i} - \mu_{2,i})^2, \quad (5)$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the distributions of real and generated samples respectively, f is the dimension of the latent space of $\mathcal{F}$, μ_1 and μ_2 are both vectors of dimension f representing the mean over all samples in the feature space of real and generated samples respectively. Σ_1 and Σ_2 are both matrices of dimension (f, f) representing the covariance matrices over all samples in the feature space of real and generated samples, respectively.

Average Pair Distance (APD). The APD metric [9] is calculated independently on each of the real and generated samples. The metric calculates the average ED between randomly selected samples in the feature space, using $\mathcal{F}$ as the latent feature extractor. The APD defines two randomly selected sets of samples, $\mathcal{S}_1$ and $\mathcal{S}_2$, of S_{apd} samples each and produces the average distance between these two sets as follows:

$$APD(\mathcal{S}, \mathcal{S}^{'}) = \frac{1}{S_{apd}} \sum_{i=1}^{S_{apd}} \sqrt{\sum_{j=1}^{f}(\mathcal{S}_{i,j} - \mathcal{S}^{'}_{i,j})^2}, \quad (6)$$

where the above is calculated on many different sets of $\mathcal{S}_1$ and $\mathcal{S}_2$ and the final presented value is the average APD found over different experiments. This is necessary to remove any bias in the selection of the two sets.

In this work, we utilized the open source software made available by [13] to calculate the above two metrics. The presented results are averaged over different initialization of the pre-trained regression model, which in our case is the deep regression model, trained only on the training set. We used a value of $S_{apd} = 20$ for the size of the randomly selected sets when calculating the APD metric.

Qualitative Analysis: Visualizing Real vs Generated. In Figure 3, we present three sequences, the first being a real sequence from the Kimore dataset, the second being generated by simply adding noise to the real sequence. The third sequence is generated using our weighted ShapeDBA approach detailed in Section 3.3. It can be seen that, visually, the noisy sample is not realistic as much as the weighted ShapeDBA generation.

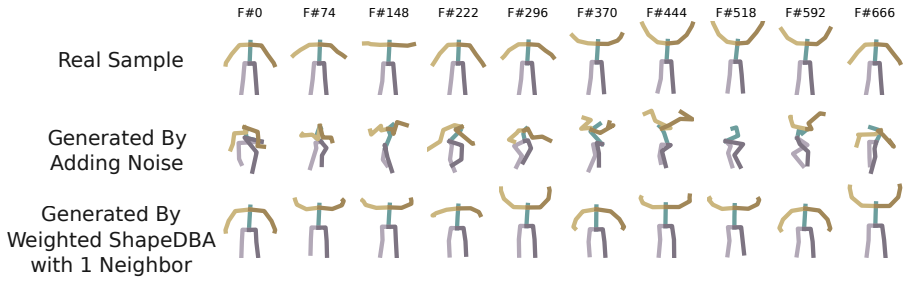

Fig. 3. Visualization of three examples: real sample (top), noisy sample (middle) and generated sample through weighted ShapeDBA (bottom). For each of the three sequences, we present 10 frames, counting from left to right.

To visualize the impact of weighted ShapeDBA, we present in Figure 4, two sequences and the produced weighted ShapeDBA following the formula in Section 3.3. It can be seen from the figure, that given the higher weight is associated to the first sequence (the reference), the motion over the temporal axis is very similar and aligned exactly with the reference. However, given a weight is assigned to the second sequence, the produced weighted ShapeDBA contains some information from the second sequence such as the height and form of the patient from whom the sequence is recorded.

Quantitative Analysis: Fidelity and Diversity. Visualizing generated sequences may produce a good enough satisfaction, however there should be a quantitative evaluation as argued in [26]. We relied on the two metrics presented in Section 4.2, the FID for fidelity evaluation and the APD for diversity evaluation. For each augmentation method, i.e., noisy and weighted ShapeDBA, we present the average FID and APD values, as well as the standard deviation. These values are aggregated over different resamples of the dataset for each exercise, as well as different initialization of the pre-trained deep regression model. In Table 1, we present the values of the FID for each exercise using the noisy augmentation method and five versions of our weighted ShapeDBA (neighbors range from 1 to 5). We also include the FID value over the real samples, which serves as a baseline, as argued in [13]. The FID of any generation method should have a close but still higher value to the computed FID on real data. This is due to the fact that generated samples cannot be more reliable, in terms of fidelity,

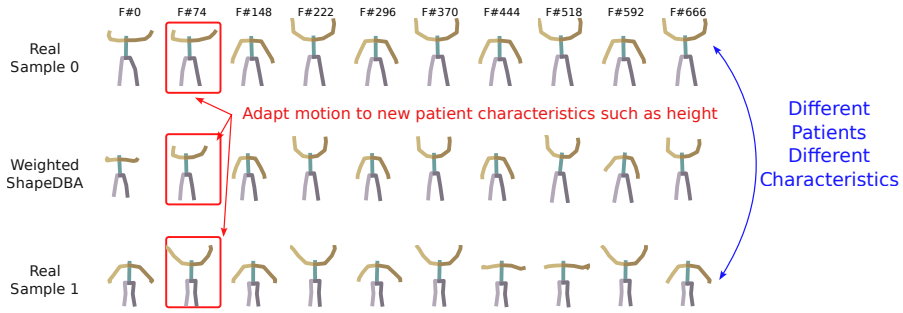

Fig. 4. Visualizing two real sequences (top and bottom) and their produced weighted ShapeDBA sequence (middle). The produced average sequence kept the temporal alignment of the exercise from the top sequence as it is associated with a higher weight than the bottom one. However the characteristic of the patient (height) is kept in the average middle sequence.

than the real set of data with itself. It can be seen from Table 1 that the FID of the noisy method is smaller than the one of the weighted ShapeDBA method. Discarding the visualization done in previous qualitative evaluation, the noisy augmentation wins over the quantitative evaluation. However from Figure 3, it can be obviously seen that the noisy augmentation method produces meaningless sequences. This is argued in [13], as there is no one *best* way to evaluate generative models, instead many approaches should be used to reach any concrete conclusion.

The same interpretation is argued for the APD values in Table 2, as the noisy method comes out as winner instead of the weighted ShapeDBA method as it has the closest value to the APD on real samples.

Table 1. The FID values of different augmentation methods and the real dataset, over different resamples of each exercise. The presented FID values include the average and standard deviation over all resamples per exercise and different initialization of the pre-trained feature extractor, **best value** second best value

	Exercise 1	**Exercise 2**	**Exercise 3**	**Exercise 4**	**Exercise 5**
Real	$02.18E^{-6} \pm 06.66E^{-7}$	$02.48E^{-6} \pm 09.82E^{-7}$	$02.19E^{-6} \pm 08.18E^{-7}$	$05.00E^{-6} \pm 02.09E^{-6}$	$03.06E^{-6} \pm 01.30E^{-6}$
Noisy	**00.07 ± 00.03**	**00.08 ± 00.03**	**00.09 ± 00.03**	**00.24 ± 00.10**	**00.07 ± 00.03**
ShapeDBA NN1	01.94 ± 00.81	04.12 ± 01.86	02.28 ± 01.10	04.19 ± 02.95	03.39 ± 01.92
ShapeDBA NN2	03.15 ± 01.06	06.62 ± 02.56	04.01 ± 02.70	05.95 ± 04.07	05.75 ± 02.34
ShapeDBA NN3	03.77 ± 01.21	07.98 ± 02.59	05.16 ± 03.39	07.16 ± 04.51	07.34 ± 02.91
ShapeDBA NN4	04.31 ± 01.24	09.38 ± 03.05	05.96 ± 03.52	08.01 ± 04.64	08.51 ± 03.76
ShapeDBA NN5	04.62 ± 01.28	10.21 ± 03.34	06.45 ± 03.69	08.90 ± 05.07	09.23 ± 04.12

4.3 Evaluation of Synthetic Sequences for Data Extension

Our second experiment evaluates the proposed approach as data extension for rehabilitation assessment. The rehabilitation assessment task is an extrinsic re-

Table 2. The APD values of different augmentation method and the real dataset, over different resamples of each exercise. The presented APD values include the average and standard deviation over all resamples per exercise, different randomly selected sets of size $S_{apd} = 20$ and different initialization of the pre-trained feature extractor.

	Exercise 1	**Exercise 2**	**Exercise 3**	**Exercise 4**	**Exercise 5**
Real	06.09 ± 00.63	06.61 ± 00.83	05.95 ± 00.78	08.24 ± 01.24	07.10 ± 01.10
Noisy	**06.05 ± 00.57**	**06.55 ± 00.79**	**05.90 ± 00.78**	**08.14 ± 01.18**	**07.00 ± 01.03**
ShapeDBA NN1	05.21 ± 00.68	05.32 ± 00.63	05.06 ± 00.68	07.14 ± 01.12	06.11 ± 01.08
ShapeDBA NN2	04.93 ± 00.69	04.96 ± 00.66	04.77 ± 00.65	06.90 ± 01.16	05.70 ± 00.98
ShapeDBA NN3	04.78 ± 00.75	04.64 ± 00.54	04.53 ± 00.66	06.84 ± 01.20	05.52 ± 00.96
ShapeDBA NN4	04.67 ± 00.74	04.47 ± 00.55	04.34 ± 00.63	06.59 ± 01.16	05.35 ± 00.91
ShapeDBA NN5	04.61 ± 00.67	04.35 ± 00.53	04.30 ± 00.61	06.50 ± 01.13	05.31 ± 00.90

gression problem, where the goal is to predict a continuous value associated to a rehabilitation sequence, i.e., the performance score.

Fully Convolutional Network for Extrinsic Regression: For the task of rehabilitation assessment, we employed a Fully Convolutional Network [34]. We used the same architecture and same hyperparameters as in [19], except for the last layer that includes a single neuron with sigmoid activation for predicting a single continuous score. A detailed view of the FCN architecture used in this work for the regression task is presented in Figure 5. The FCN architecture consists of three stacked convolution blocks followed by an aggregation layer, Global Average Pooling, before being fed to a Fully Connected layer to predict one single value. Each convolution block consists on a one dimensional convolution layer, a Batch Normalization layer in order to re-scale the features to the same range of values, and a ReLU activation in order to detect only the activated features of each convolution filter.

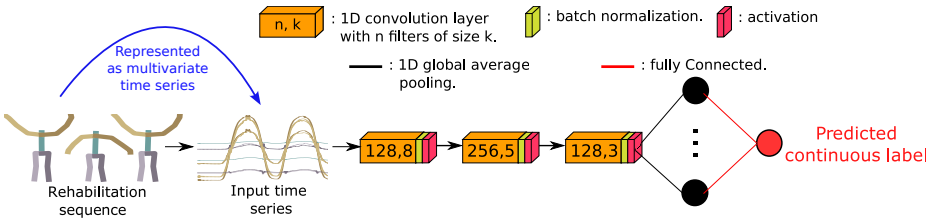

Fig. 5. The Fully Convolutional Network (FCN) used for the task of extrinsic regression on rehabilitation sequences represented as multivariate time series.

Regression Metrics: For comparing the score predicted by the considered models with the clinical score provided by experts, we followed [2] and used two metrics: the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE). Given two sets of N scores, $\mathbf{y}$ as the ground truth and $\hat{\mathbf{y}}$ as the

predictions, both the MAE and RMSE are computed as follows:

$$MAE(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|, \tag{7}$$

$$RMSE(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}. \tag{8}$$

Quantitative Analysis: We evaluated the performance of the considered FCN model for extrinsic regression, while trained on the different sets of rehabilitation sequences described in Section 4.1, concatenated with the *reference set*. Average MAE and RMSE errors (± std) are reported in Table 3. We can first notice that, in most of the cases, a better performance is obtained when noisy sequences are added to the training set (*Ref. + Noise*), with respect to the case where only the original training set is used (*Ref.*). This shows the FCN model overfits the training set and considering data extension allows slightly improving generalization.

Moreover, it can be seen from Table 3 that the best error values for both metrics are obtained when the FCN model is trained on the rehabilitation set extended with ShapeDBA averages. This shows that the synthetic average sequences not only allow preventing overfitting but also correspond to more realistic rehabilitation motion so that a FCN model can better learn variations of rehabilitation exercises.

Finally, we can observe that, for three exercises among five, the best performance is obtained when training data includes average sequences computed using a single neighbor (*Ref. + ShapeDBA NN1*). This suggests that considering a larger neighborhood may result in incoherent average sequences in some cases as neighbors are not lying in a continuous subspace.

Table 3. MAE and RMSE errors obtained for all compared approaches on each exercise separately. Best values are emphasized in bold, while second best values are underlined.

Training Set	Exercise 1	Exercise 2	Exercise 3	Exercise 4	Exercise 5
			MAE		
Ref.	0.206 ± 0.069	0.202 ± 0.037	0.204 ± 0.055	0.184 ± 0.068	0.224 ± 0.058
Ref. + Noise	0.186 ± 0.065	0.172 ± 0.040	0.203 ± 0.045	0.185 ± 0.073	0.229 ± 0.069
Ref. + ShapeDBA NN1	0.167 ± 0.070	**0.175 ± 0.030**	**0.182 ± 0.051**	**0.141 ± 0.062**	**0.208 ± 0.079**
Ref. + ShapeDBA NN2	0.169 ± 0.057	0.177 ± 0.041	0.194 ± 0.041	0.168 ± 0.056	0.226 ± 0.066
Ref. + ShapeDBA NN3	0.173 ± 0.063	0.183 ± 0.047	0.199 ± 0.058	0.168 ± 0.083	0.225 ± 0.055
Ref. + ShapeDBA NN4	**0.168 ± 0.059**	0.179 ± 0.043	0.199 ± 0.043	0.180 ± 0.080	0.231 ± 0.060
Ref. + ShapeDBA NN5	0.166 ± 0.067	0.185 ± 0.043	0.201 ± 0.050	0.182 ± 0.089	0.226 ± 0.061
			RMSE		
Ref.	0.251 ± 0.083	0.247 ± 0.045	0.248 ± 0.065	0.230 ± 0.083	0.267 ± 0.073
Ref. + Noise	0.203 ± 0.078	0.226 ± 0.043	0.238 ± 0.046	0.227 ± 0.090	0.274 ± 0.092
Ref. + ShapeDBA NN1	0.199 ± 0.087	**0.226 ± 0.036**	**0.214 ± 0.054**	**0.178 ± 0.074**	0.251 ± 0.094
Ref. + ShapeDBA NN2	0.203 ± 0.075	0.232 ± 0.052	0.226 ± 0.044	0.210 ± 0.074	0.268 ± 0.083
Ref. + ShapeDBA NN3	0.205 ± 0.082	0.235 ± 0.050	0.240 ± 0.062	0.214 ± 0.105	**0.268 ± 0.066**
Ref. + ShapeDBA NN4	**0.198 ± 0.071**	0.235 ± 0.050	0.234 ± 0.048	0.230 ± 0.105	0.279 ± 0.070
Ref. + ShapeDBA NN5	0.202 ± 0.079	0.230 ± 0.049	0.244 ± 0.057	0.231 ± 0.109	0.280 ± 0.080

5 Conclusion

In this work, we addressed the problem of rehabilitation assessment with limited data by employing a state-of-the-art elastic averaging method, ShapeDBA, to extend the amount of available training data. Our approach leverages weighted averaging to generate synthetic rehabilitation sequences and associate new continuous scores with them, thus enhancing the assessment process. Experimental evaluation demonstrated the promise of this method to generate coherent rehabilitation sequences.

However, our method remains dependent on the real data distribution, particularly for continuous labels, which limits control over the generated sequences and scores. Future work will focus on evaluating this approach on more diverse datasets and models as well as investigating how to improve control over synthetic sequences, ultimately aiming to enhance data-driven rehabilitation assessment and patient outcomes.

Acknowledgment

This work was supported by the ANR DELEGATION project (grant ANR-21-CE23-0014) of the French Agence Nationale de la Recherche. The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data. The authors would also like to thank the creators and providers of the Kimore dataset.

References

1. Blanchard, A., Nguyen, S.M., Devanne, M., Simonnet, M., Le Goff-Pronost, M., Rémy-Néris, O.: Technical feasibility of supervision of stretching exercises by a humanoid robot coach for chronic low back pain: The r-cool randomized trial. BioMed Research International **2022**(1), 1–10 (2022), https://doi.org/10.1155/2022/5667223
2. Capecci, M., Ceravolo, M.G., Ferracuti, F., Iarlori, S., Kyrki, V., Monteriu, A., Romeo, L., Verdini, F.: A hidden semi-markov model based approach for rehabilitation exercise assessment. Journal of biomedical informatics **78**, 1–11 (2018)
3. Capecci, M., Ceravolo, M.G., Ferracuti, F., Iarlori, S., Monteriu, A., Romeo, L., Verdini, F.: The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation. IEEE Transactions on Neural Systems and Rehabilitation Engineering **27**(7), 1436–1448 (2019)
4. Chen, J., Yang, W., Liu, C., Yao, L.: A data augmentation method for skeleton-based action recognition with relative features. Applied Sciences **11**(23), 11481 (2021)
5. Deb, S., Islam, M.F., Rahman, S., Rahman, S.: Graph convolutional networks for assessment of physical rehabilitation exercises. IEEE Transactions on Neural Systems and Rehabilitation Engineering **30**, 410–419 (2022)

6. Devanne, M., et al.: Multi-level motion analysis for physical exercises assessment in kinaesthetic rehabilitation. In: 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids). pp. 529–534. IEEE (2017)
7. Forestier, G., Petitjean, F., Webb, G., Dau, H.A., Keogh, E.: Generating synthetic time series to augment sparse datasets. In: IEEE International Conference on Data Mining (ICDM). pp. 865–870 (2017), https://doi.org/10.1109/ICDM.2017.106
8. Guijo-Rubio, D., Middlehurst, M., Arcencio, G., Silva, D.F., Bagnall, A.: Unsupervised feature based algorithms for time series extrinsic regression. Data Mining and Knowledge Discovery pp. 1–45 (2024)
9. Guo, C., Zuo, X., Wang, S., Zou, S., Sun, Q., Deng, A., Gong, M., Cheng, L.: Action2motion: Conditioned generation of 3d human motions. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2021–2029 (2020)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Huynh-The, T., Hua, C.H., Kim, D.S.: Encoding pose features to images with data augmentation for 3-d action recognition. IEEE Transactions on Industrial Informatics **16**(5), 3100–3111 (2019)
12. Ismail-Fawaz, A., Devanne, M., Berretti, S., Weber, J., Forestier, G.: Lite: Light inception with boosting techniques for time series classification. In: 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA). pp. 1–10. IEEE (2023)
13. Ismail-Fawaz, A., Devanne, M., Berretti, S., Weber, J., Forestier, G.: Establishing a unified evaluation framework for human motion generation: A comparative analysis of metrics. arXiv preprint arXiv:2405.07680 (2024)
14. Ismail-Fawaz, A., Devanne, M., Berretti, S., Weber, J., Forestier, G.: A supervised variational auto-encoder for human motion generation using convolutional neural networks. In: 4th International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI) (2024)
15. Ismail-Fawaz, A., Devanne, M., Berretti, S., Weber, J., Forestier, G.: Weighted elastic barycetner averaging to augment time series data. https://github.com/MSD-IRIMAS/Augmenting-TSC-Elastic-Averaging (2024)
16. Ismail-Fawaz, A., Devanne, M., Weber, J., Forestier, G.: Deep learning for time series classification using new hand-crafted convolution filters. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 972–981. IEEE (2022)
17. Ismail-Fawaz, A., Ismail Fawaz, H., Petitjean, F., Devanne, M., Weber, J., Berretti, S., Webb, G.I., Forestier, G.: Shapedba: Generating effective time series prototypes using shapedtw barycenter averaging. In: International Workshop on Advanced Analytics and Learning on Temporal Data. pp. 127–142. Springer (2023)
18. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Data augmentation using synthetic data for time series classification with deep residual networks. In: ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data (2018)
19. Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data mining and knowledge discovery **33**(4), 917–963 (2019)
20. Ismayilzada, E., Devanne, M., , Weber, J., Forestier, G.: Time series extrinsic regression for physical rehabilitation assessment. In: Upper Rhine Artificial Intelligence Symposium (URAI) (2023), undefined

21. Lee, I., Kim, D., Kang, S., Lee, S.: Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1012–1020 (2017)
22. Liao, Y., Vakanski, A., Xian, M.: A deep learning framework for assessing physical rehabilitation exercises. IEEE Transactions on Neural Systems and Rehabilitation Engineering **28**(2), 468–477 (2020)
23. Middlehurst, M., Ismail-Fawaz, A., Guillaume, A., Holder, C., Rubio, D.G., Bulatova, G., Tsaprounis, L., Mentel, L., Walter, M., Schäfer, P., et al.: aeon: a python toolkit for learning from time series. arXiv preprint arXiv:2406.14231 (2024)
24. Mourchid, Y., Slama, R.: Mr-stgn: Multi-residual spatio temporal graph network using attention fusion for patient action assessment. In: 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP). pp. 1–6. IEEE (2023)
25. Müller, M.: Dynamic time warping. Information retrieval for music and motion pp. 69–84 (2007)
26. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning. pp. 7176–7185. PMLR (2020)
27. Nguyen, S., Devanne, M., Remy Neris, O., Lempereur, M., Thepaut, A.: A medical low-back pain physical rehabilitation database for human body movement analysis. In: International Joint Conference on Neural Networks (IJCNN). IEEE (2024)
28. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern recognition **44**(3), 678–693 (2011)
29. Petrovich, M., Black, M.J., Varol, G.: Action-conditioned 3d human motion synthesis with transformer vae. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10985–10995 (2021)
30. Pialla, G., Devanne, M., Weber, J., Idoumghar, L., Forestier, G.: Data augmentation for time series classification with deep learning models. In: International Workshop on Advanced Analytics and Learning on Temporal Data (2022)
31. Rao, H., Xu, S., Hu, X., Cheng, J., Hu, B.: Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. Information Sciences **569**, 90–109 (2021)
32. Vakanski, A., Jun, H.p., Paul, D., Baker, R.: A data set of human body movements for physical rehabilitation exercises. Data **3**(1), 2 (2018)
33. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)
34. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International joint conference on neural networks (IJCNN). pp. 1578–1585. IEEE (2017)
35. Xin, C., Kim, S., Cho, Y., Park, K.S.: Enhancing human action recognition with 3d skeleton data: A comprehensive study of deep learning and data augmentation. Electronics **13**(4), 747 (2024)
36. Zhao, J., Itti, L.: shapedtw: Shape dynamic time warping. Pattern Recognition **74**, 171–184 (2018)