Short: Deep Learning Approach to Skeletal Performance Evaluation of Physical Therapy Exercises

Bhanu Garg  
University of California, San Diego  
San Diego, California, USA  
b garg@ucsd.edu

Pamela Cosman  
University of California, San Diego  
San Diego, California, USA  
pcosman@eng.ucsd.edu

Alexander Postlmayr  
University of California, San Diego  
San Diego, California, USA  
apostlmaya@ucsd.edu

Sujit Dey  
University of California, San Diego  
San Diego, California, USA  
de y@eng.ucsd.edu

ABSTRACT
At-home exercising strongly predicts physical therapy patient outcomes, underscoring the need for analyzing patient behaviors at home via remote patient monitoring. Contemporary methods for remote patient monitoring rely on specialized sensors, i.e., Inertial Measurement Units, RGB-Depth cameras, motion capture systems, or stereo vision which are costly and not scalable to all physical therapy patients. Here, we observe a lack of literature using only a monocular RGB camera. In this paper, we demonstrate a skeletal feedback model for at-home exercises using only video acquired from a smartphone camera. We propose models for (i) Patient Performance Evaluation - which classifies the correctness of exercises, and (ii) Guidance - which identifies why the exercise went wrong so the patient can correct themselves. We use these models on our dataset of four common physical therapy exercises labeled by a physical therapist. Our results demonstrate the feasibility of using skeletal data from state-of-the-art 3D human pose estimation models for physical rehabilitation exercise evaluation and guidance. Thus, we enable remote patient monitoring and guidance from a single camera - making it highly cost-effective and scalable.

KEYWORDS
PT rehabilitation, Patient Performance Evaluation, posture correction, deep learning, self attention, home monitoring

1 INTRODUCTION
The field of human action evaluation (HAE) is broad in its applications, ranging from gait analysis [27] to judging Olympic performance [22]. Physical rehabilitation is one of many applications which can significantly benefit from using HAE technologies [3]. Using skeleton features in HAE was shown to be promising in [9]; however, these studies are limited to low fidelity exercises, i.e., large amplitude movements.

Using specialized sensors, such as Inertial Measurement Units (IMUs) [17], RGB-D cameras [14],[19],[31],[20], or motion capture systems [6] for HAE have shown promising results in displaying accurate assessment and quantification of rehabilitation and strength exercises. While these specialized sensor-based assessment technologies provide high accuracy, they are limited in applicability due to the inherent cost and complex nature of obtaining specialized hardware for action evaluation. Moreover, standardizing skeleton data and deep feature representation methods from the sensors is another key issue in developing reliable quality assessment algorithms for HAE [12]. Thus, relying on specialized hardware limits practical applications of HAE and prevents the creation of standard datasets required to advance the technology.

Recent advances in 3D Human Pose Estimation (HPE) such as [21], [16] have allowed for the feature extraction of skeletal key points from a monocular RGB camera. These 3D HPEs are popularly used to predict key joint positions on the body [11]. In this paper, we show that classical machine learning methods such as Dynamic Time Warping (DTW) are limited for monocular RGB HAE due to the inherent noise associated with predicting the depth dimension from a monocular RGB camera. Using 2D skeletal information in conjunction with deep learning has shown promising results for regression analysis [13]. In HAE for physical therapy (PT), getting 3D skeletons is necessary to clinically assess range of motion in key joints. Furthermore, physical rehabilitation requires corrective feedback for improving patient outcomes [8] [10]. Although monocular RGB 3D HPE is promising, we observe no HAE with high-fidelity feedback in the literature.

In our work, we propose a framework for skeleton-based HAE for PT exercises from a single camera, that evaluates patient repetitions as correct or not, and offers explainability for correcting the incorrect movements. For this paper, we narrow the scope to human Patient Performance Evaluation and guidance, assuming 3D skeletal features are obtained from a 3D human pose estimator.
2 RELATED WORK

We review various approaches proposed for general HAE and for PT applications, and highlight the major differences between them and our work. Although specialized sensor-based systems for physical rehabilitation are popular in the literature [15], we observe limited approaches based on a single RGB camera. To our knowledge, there are no 3D skeletal feature, monocular-based approaches. The literature highlights the importance of 3D skeletal features as a constituent to HAE in physical therapy; therefore, our work complements existing work that quantifies physiotherapy metrics such as range-of-motion and joint angle-based success criteria implicitly, by learning from data.

2.1 HAE based on complex sensors

Rooted in classical signal processing [4], matching techniques have seen some success in RGB-D applications [29]. Another classical approach [7] was used as a screening tool by performing anomaly detection on several activities of daily living (ADL). This system allows for health service provider intervention for given neuromusculoskeletal conditions but does not address rehabilitation. Moreover, we observe classical analytical methods are insufficient for datasets with low signal-to-noise ratio, i.e., collected using monocular RGB with skeletal features engineered using state-of-the-art HPE models.

Deep learning approaches have been popular for regression analysis [24] [30] on RGB-D, optical tracking system [14], and IMU [26] datasets. These regression models do not provide any patient feedback. The ability to use deep learning to diagnose and track the progression of Alzheimer’s Disease was shown in [31].

2.2 HAE based on a single RGB camera

Outside of physical therapy, approaches based on a single RGB camera for HAE have been shown by [13], which utilized 2D skeletal features extracted from monocular RGB to train deep learning regression models on the UNLV Olympic Diving and MIT Olympic Ice Skating Scoring datasets. Pseudo3D was used to extract spatiotemporal features from video on the UNLV Diving dataset [5]. The feature engineering here [23] is limited to pseudo representations of human pose estimation.

For physical rehabilitation, [25] uses a regression-based quantitative scoring model trained on the KIMORE dataset using monocular RGB. A lack of accurate determination of joint angles from 2D skeleton data limits the extent of evaluation of exercises. For example, indicators based on range of motion of joints, a prevalent metric used in PT, cannot be modeled (even implicitly) by 2D joint data. Feedback on static exercises only was shown in [18]; these comprise a small fraction of rehabilitative exercises.

3 METHODOLOGY

We begin by defining key terms and notation in Section 3.1. In Section 3.2, we explain our data collection method, while in Section 3.3, we build the case to show that classical methods such as Dynamic Time Warping are not suitable for evaluating repetitions and providing guidance results for data collected using monocular RGB camera videos. We then describe our proposed methods in Sections 3.4 and 3.5.

Figure 1: Process Flow: (A) Patient records a video performing an exercise, (B) each frame is skeletonized, (C) the skeletons are transformed to an angular domain and post-processed, (D) the resulting vector is fed to PPE and Guidance models, (E) patient receives feedback.

3.1 Terminology

We focus on HAE using a monocular RGB camera which is available in most smartphones and tablet computers. For simplicity, any data is from a monocular RGB camera unless otherwise specified. For Patient Performance Evaluation (PPE), we formulate and define the following tasks in our processing and data pipeline.

- **Skeletonization**: Given an RGB frame containing a human, skeletonization refers to the process of extracting the 3D (x,y,z) coordinates of key joint positions and constructing a stick figure-like model of the human. The term **skeleton** refers to the constructed stick figure. This paper uses a custom skeletonizer that gives results similar to [21].

- **Segmentation**: Given a sequence of skeletons from the video feed, segmentation is the process of extracting exercise repetitions from the entire exercise session. Each exercise repetition has start and stop frame numbers in the video. We use r to denote a repetition.

- **Patient Performance Evaluation (PPE)**: Given an exercise repetition, PPE assigns the exercise repetition a label from the set \{good, bad\}. Ground truth labels are determined by an expert, i.e., a licensed physical therapist.

- **Guidance**: For a repetition that has been tagged bad, guidance is the suggestion to correct the form of that repetition. For example, while doing a deadlift, possible guidance could be “keep back straight” if the patient has a rounded back.

Fig. 1 shows the overall flow of our proposed process. Given the abundance of literature on human pose estimation and segmentation, we narrow our attention to PPE and guidance methods.

3.2 Feature Engineering

We record student volunteers performing exercises on phone cameras. We provide more details on data collection in Section 4.1. We use a custom skeletonization model, which outputs the 3D positions of 17 key joints as in [21]. The videos are then skeletonized and segmented, and the individual repetitions are fed as input to the subsequent steps.

We begin by selecting joint angles relevant to the exercise \(E\) as determined by an expert. We note that the selected joint angles \(a_j\) could be those that move during the exercise as well as those that
might be required to stay stationary. Let \( \mathcal{A}_E = \{a_j\} \) be the set of such selected joint angles for exercise \( E \).

To formalize, for a repetition \( r \) belonging to exercise \( E \) of time length \( T \) video frames, we skeletonize to obtain a tensor of shape \((T, 17, 3)\), for 17 joints and 3 spatial dimensions. For each frame \( i \in [1, \ldots, T] \), we compute the angles for the joints in \( \mathcal{A}_E \). Let \( v_{(j,i)} \) be the angle for joint \( a_j \) at time \( i \). The 2D tensor \( V = [v_{(j,i)}] \) of dimensions \((|\mathcal{A}_E|, T)\) is then smoothed and subsampled as described in the paragraph below to get \( \tilde{V} \). This \( \tilde{V} \) is used as an input to the models described in Sections 3.4.1 and 3.4.2.

**Smoothing filter:** We use an averaging filter with window size 5 on the time series of individual joint angles to smooth out outliers.

**Subsampling:** As different subjects perform exercise repetitions at different rates, the \( V \) vary in width. This can cause problems in training deep neural networks including difficulties in batch training and data preprocessing. Based on our discussions with the physical therapist, the exercise rate is seldom useful in classifying a repetition as good or bad. More importance is given to the form, which does not depend on the rate. Hence, we subsample each repetition to 20 equidistant frames.

### 3.3 Classical Methods: Dynamic Time Warping

Dynamic Time Warping (DTW) is a method for measuring similarity between two temporal sequences. It was used successfully in the context of physical therapy exercises in [29] where data was collected with a Microsoft Kinect Camera. In our experiments, DTW was collected with a Microsoft Kinect Camera. In our experiments, DTW could be applied to patient repetitions to compare them with a "ground truth" coming from our physical therapist consultant. In concept, it could allow the cumulative difference across all key angles to be used as a general threshold for PPE while using angle-specific thresholds to provide feedback on any particularly incorrect angles. Despite its efficacy with RGB-D based datasets, we observed poor results on our 3D HPE skeletal data. Fig. 2 plots the distribution of similarity scores per joint angle from DTW for good and bad repetitions; the substantial overlap between the two distributions makes the thresholding-based approach infeasible. Thus, we turn to deep learning-based methods that are more potent in finding fine patterns in the data.

### 3.4 Patient Performance Evaluation (PPE)

In this section, we describe our neural network-based approach to PPE. We first describe the data processing steps and then the model architectures.

#### 3.4.1 CNN-based PPE

In this section, we describe the convolutional neural network (CNN) architecture used for classifying \( \tilde{V} \) to the PPE label set \( Y = \{\text{good, bad}\} \). In the first layer, we apply a convolution filter on the time axis (temporal convolution), i.e., a kernel of shape \((3, 1)\) and output channels 32. Then the next layer is convolution on the angles (spatial convolution), i.e., a kernel of shape \((1, |\mathcal{A}_E|)\) and output channels 64. After each of these two convolutions, a ReLU non-linearity is followed by a batch norm. Finally, the feature map from the last convolution is average-pooled, followed by a binary classification head. The intuition for choosing the above architecture is that the temporal convolution learns the temporal context at time \( t_i \) of the angle \( a_j \). Once the context of the angle has been set, the model learns the interdependence relations among the joint angles. To the best of our knowledge, ours is the first method to view human Patient Performance Evaluation as a Spatio-Temporal convolution network.

#### 3.4.2 Attention based PPE

In this section, we describe the architecture of the attention-based transformer classifier used for PPE. Attention networks have gained popularity by learning to attend to values at different timestamps and have been successful in natural language processing and computer vision [28]. The number of heads is chosen to be 2, with a 512 embedding dimension for the encoder. A classification head follows the encoding head.

### 3.5 Guidance and Exercise Feedback

This section proposes our method for generating guidance and feedback mechanisms for the bad repetitions. An autoencoder is a model that encodes the input into a feature map and then uses the encoding to decode it back to the input. It is used to train an
underlying model distribution of the data. With this context, we first train an autoencoder using only the good repetitions to make the neural network learn a robust good template of the exercise.

The mean and variance (μaj, σaj) of the reconstruction error for each joint angle are collected. Then for a given bad repetition, the reconstruction error ̂eaj from the autoencoder is computed. The angles with the highest z-score: zaj = ̂eaj−μaj/σaj are selected as the ones responsible for the repetition being bad. The process is shown visually in Fig. 4.

AutoEncoder: Following similar arguments on the efficacy of attention-based models on time series data as described in Section 3.4, we propose to use an attention based autoencoder. While implementing a corresponding CNN-based auto-encoder, we found that maintaining the convolutional upsampling layers to consistently output the same size vector as the input for different exercises non-trivial, and adding unnecessary complexity to our model pipeline. In the attention-based autoencoder model, we use the default number of dimensions for the embeddings, with the number of heads equal to |ÅE|, and mean square error (MSE) as the loss function.

4 EXPERIMENTS

4.1 Data

In this section, we discuss the details of our dataset, which we use to demonstrate the efficacy of our proposed PPE and guidance models. As existing datasets [14] [2] [1] do not have corrective feedback and classification scores, we were required to create our own.

4.1.1 General Setup. Based on consultation with a physical therapist, we selected four exercises that are frequently prescribed in physical therapy, involve compound movements, and collectively cover multiple focus areas (shoulder, legs, hips). The exercises are:

- double leg Romanian dead-lift (DoubleRDL)
- single leg Romanian dead-lift (SingleRDL)
- single leg mini squat (SingleMS)
- rotator cuff (RotatorCuff)

Students working on the project volunteered to be recorded performing exercises. We selected 10 subjects for our data collection. Before recording each exercise, the subjects were shown a demonstration video created by the physical therapist that shows several repetitions of the exercise with proper form. The subjects were asked to recreate 10 repetitions seen in the video to the best of their ability. For unilateral exercises, subjects were asked to perform 5 repetitions on each side. After the first 10 repetitions, subjects were informed about common mistakes seen in physical rehabilitation. For example, a DoubleRDL is often performed with the subject’s spine being too rounded with too much scapular protraction or the subject’s knees being too bent or locked out. These subtle mistakes have a profound impact on muscle activation during the exercise. The subjects were instructed to incorporate these incorrect postures into an additional 10 repetitions.

To test our proposed methods’ robustness to different viewing angles and camera types, we recorded subjects using five different smartphones (iPhone Models X, 11, 7, 8, and Google Pixel 3A) placed on tripods approximately 4 feet high and at five different angles (30, 60, 90, 120, and 150 degrees) to the subject. To summarize, data consists of 1000 repetitions for each exercise from 10 subjects, and 5 camera angles.

4.1.2 Labeling. For each video recording, each exercise was segmented manually into repetitions which were classified as being good or bad, and the two most erroneous movements that needed correction were noted. Our final dataset includes 15% good repetitions for DoubleRDL, 15% for SingleRDL, 42% for SingleMS, and 50% for RotatorCuff. This variation is due to the fact that subjects were more knowledgeable about proper exercise form for some exercises compared to others.

4.2 Results

We split the 10 subjects into 3 folds (with 4, 3, and 3 subjects), and train the model on 2 of the folds, and evaluate the model on the remaining fold. We show the mean F1 score of the 3-fold cross validation.

Patient Performance Evaluation For PPE, we use negative log likelihood loss for training. We use the Adam optimizer for training with learning rate 1e−4 and default weight parameters β = (0.9, 0.98), and batch size 16. We report the F1 score of the binary classification model. We see that both the CNN and Attention-based models consistently outperform DTW (see Table 1). Between the CNN and Attention-based models, the Attention-based model performs slightly better than the CNN-based model.

Our results show promise that our novel solutions can classify exercises based on single camera input.

Guidance We map each joint angle to an action item that the subject could do to correct that angle based on the output of the autoencoder. For guidance models, we report the top − 2 accuracy. Specifically, it is considered correct if either of the two most erroneous predicted angles match the PT judgment. Human body movements have constraints and the joint angles do not operate in isolation from one another. Correcting one angle could affect the correctness of the other angle - for example, in DoubleRDL, making the back straight would still be conducive to correcting locked knees. Hence, the top-2 accuracy metric is justified.

For training the Attention autoencoder, we use the Adam optimizer with learning rate 1e−3 and default weight parameters β = (0.9, 0.98) with batch size 32. In Table 1, we see that the Attention-based autoencoder method outperforms DTW. Ours is the first deep learning-based approach to PT guidance.

5 CONCLUSION AND FUTURE WORK

We propose deep-learning Patient Performance Evaluation and guidance models for PT rehabilitation and at-home monitoring using only a smartphone camera. Our CNN and Attention networks
show improved results on PPE and guidance for all four exercises. We observe the importance of a robust dataset as a priority for future work. In contrast, DTW would not benefit from such a dataset. To the best of our knowledge, this is the first deep learning-based approach for PT guidance. In the future, we plan to run clinical trials on the effectiveness of our method on patient outcomes and expand our exercise library to many more exercises.

REFERENCES


Exercise | Guidance Criteria
--- | ---
DoubleRDL | Knees Too Bent, Knees Locked, Back Too Round, Feet Too Far Apart
SingleRDL | Knees Too Bent, Knee Locked, Back Too Round, Leg Not In-Line
SingleMS | Hips Not Level, Squat Too Low, Twisting Torso
RotatorCuff | Twisting Torso, Arm Too Extended, Lifting Arm Too High/Low

Table 2: Guidance cues defined by PT